Genome-wide transposon analyses: annotation, movement and impact on plant function and evolution

Elizabeth Marie Hénaff

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA

PROGRAMA DE BIOTECNOLOGIA VEGETAL

Genome-wide transposon analyses: annotation, movement and impact on plant function and evolution

Memòria presentada per Elizabeth Marie Hénaff per optar al títol de doctor per la universitat de Barcelona

Director
Josep Maria Casacuberta Suñer

Doctoranda
Elizabeth Marie Hénaff

Tutor
Albert Ferrer Prats

2013

Index

Introduction

INTRODUCTION

The diversity of life forms around us is astounding: a walk in the woods, or even down the street, shows us organisms of different morphologies: two legs, four legs, wings; different capacity of interaction with our environment: plants photosynthesizing while bacteria break down our garbage; even one neighbor's german shepherd and the other's chihuahua seem to have little in common besides the name of dog. How can life take on so many forms? Our understanding of the plants and animals that surround us – and ourselves – has been constantly progressing from the macroscopic to the microscopic, and smaller. Darwin postulated the laws of inheritance and selection two centuries ago, and in the past century we started to understand their molecular basis and the existence of genes as basic units of function. Genes … the answer is in the genes! If a gene codes for a protein, and a protein for a function, then the more complex the functions, the more genes we have... no?

Wrong! The number of genes is largely conserved across organisms, even though their complexity varies hugely. It was indeed quite a blow to our self-esteem when we realized the worm *C. elegans* has almost just as many genes as humans (Cowley and Oakey 2013). While there is some increase of genes when comparing the most simple eukaryotes to the most complex ones it is clear that organism complexity is not the result of the number of genes (**Table 1**).

genome	size (Mbp)	number of genes	reference
Saccharomyces cerevisiae	11.7	5700	www.broadinstitute.org
Drosophila melanogaster	120	15200	Adams et al (2000)
Caenorhabditis elegans	97	19000	CSC (1998)
Homo sapiens	2850	24000	IHGSC (2004)
Arabidopsis thaliana	120	26200	AGI (2000)
Mus musculus	2500	30000	MGSC (2002)

CSC: C. elegans sequencing consortium
AGI: Arabidopsis genome initiative
IHGSC: International human genome sequencing consortium
MGSC: Mouse genome sequencing consortium

Table 1: genome size and number of genes in various sequenced genomes

Therefore is has been postulated that the complexity of an organism arises from the complexity of its gene regulation, rather than the number of genes. This regulation must come then from the non-gene part of the genome. The analysis of individual genes has allowed the characterization of regulatory elements, such as the basal promoter, activators and repressors that modify transcription, the terminator that controls transcript termination, as well as splicing factors that allow the processing of the transcript into its mature form that can also be regulated. However, up to now, most of these elements have been associated to single genes and have been described close to them, being integrated into the notion of gene itself, which now defines the coding sequences together with their proximal regulatory elements.

We now know that genes constitute but a small portion of genomes, (about 5% of the human genome (Venter et al. 2001)) and the rest of it has been bundled into the catch-all term of “non-coding sequences” or, for the more pessimistic, “junk”. But as in any attic there are always gems to be found in the junk, and maybe there amidst the dusty old bits and bobs lie the answers to our past, and a wealth of possibilities for the future. Barbara McClintock, in her seminal experiments on maize chromosome breakage (McClintock 1983), and in her visionary and imaginative interpretation of the results, postulated the existence within this junk DNA of “controlling elements” which, by their movement, influence gene expression.

The advent of whole-genome sequencing has enabled us to get a more complete picture of what is in a genome, and with that has come the surprise that a significant part of all genomes characterized is constituted of transposable elements (TEs). These jumping genes have enjoyed a turnaround from being parasitic genomic junk to recognized as potent drivers of evolution. In this introduction I will give a general overview of what are transposable elements, and give some examples of how individual elements were discovered and discuss how next generation sequencing is changing our approach and scope of understanding. Then I will expose some of the many manners in which TEs can and have influenced the host genomes they occupy, and lay out the goals of this study.

What is a transposable element?

Transposable elements are mobile genetic sequences, meaning that they have the capacity to change their position within the genome of a single cell. There are two main categories of transposons, that differ by their means of transposition. Class I transposons, or retrotransposons, transpose via an RNA intermediate, leaving a copy behind and introducing a new copy in a different genomic location. These are the “copy-paste” transposons. Class II elements, or DNA transposons, are excised from their position and integrated into a different genomic location, hence are dubbed the “cut and paste” kind. Elements of both classes can be found in families of similar elements, though retrotransposons tend to form larger families due to their replicative nature. Within each class of transposon, one can categorize them further into superfamilies according to broad features such as the structure of encoded proteins or non-coding regions, or target site duplication (TSD) length (**Figure 1**).

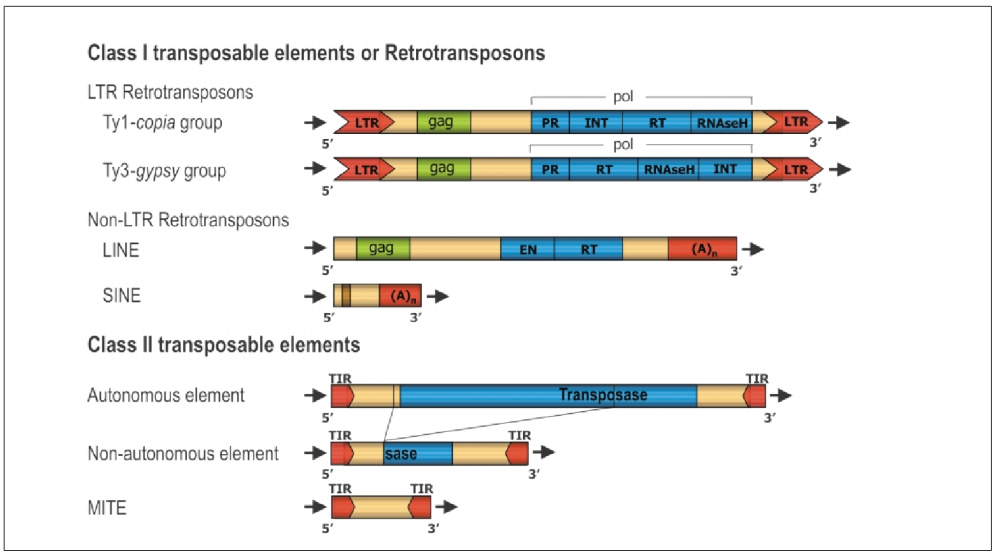


Figure 1: classes and major superfamilies of TEs.

Taken with permission from Casacuberta and Santiago 2003

Within Class I retrotransposons, the main distinction is between LTR retrotransposons and all the others, bagged into the term non-LTR retrotransposons. The former are flanked by characteristic LTRs, or Long Terminal Repeats, (100bp to several Kb in length) and encode a set of proteins. The promoters for expression of these genes are within the LTR, and therefore these elements have the particularity of containing two promoter sets, one in either identical LTR, conferring them a potential effect of read-out transcription from their 3' LTR. The genes encoded by these elements include those necessary for their mode of transposition: GAG and POL, synthesized as a polyprotein. GAG forms a virus-like particle (VLP), which packages the transposon RNA, reverse transcriptase (RT), and integrase (INT). Within the VLP the two RNA molecules are used by the reverse transcriptase to produce a cDNA copy of the retrotransposon. The VLP then translocates its cargo to the nucleus where the integrase cuts the genomic DNA molecule and inserts the newly retrotranscribed copy of the retrotransposon (**Figure 2**).

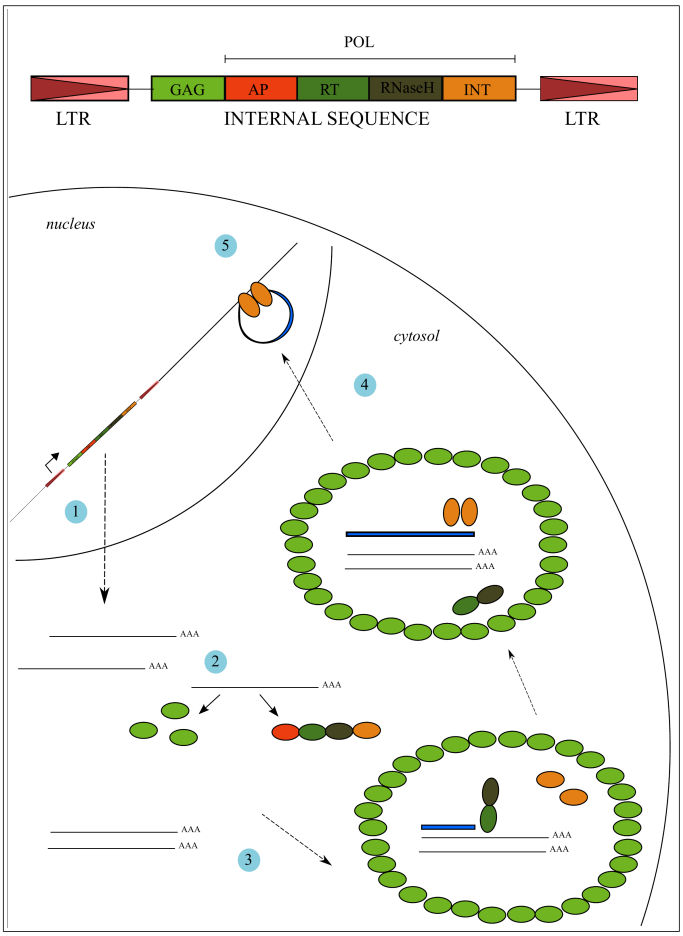


Figure 2: life cycle of a retrotransposon.

The order of proteins in the POL polyprotein are that of a gypsy type element. The proteins are GAG: capsid protein; AP: aspartic proteinase; RT: reverse transcriptase; RNAseH and INT: integrase. The steps are:

1) *transcription from the promoter region in the LTR*
2) transcription of GAG and the polyprotein, the cleaved by AP
3) two mRNAs, RT-RNAseH and INT are packaged in a Virus-Like-Particle
4) retrotranscription of the RNA by RT
5) localization of the VLP to the nucleus and passage of the RNA-INT into the nucleus
6) integration of the cDNA into the genome.

Some LTR retrotransposons also encode an envelope-like protein similar to that of retroviruses, which shows that these two mobile elements have a common ancestral origin. Whether retroviruses are old retrotransposons that acquired the capacity to leave the cell or retrotransposons are old retroviruses that lost it is not clear. LTR retrotransposons are classified into two major superfamilies, differentiated by the order of subunits in the POL polyprotein: in the *copia* superfamily INT precedes RT and RNAseH, while in *gypsy* type elements INT is found last. LTR retrotransposons are particularly abundant in plants, where they can constitute a large portion if the transposon fraction of the genome.

LINEs are the most common non-LTR retrotransposons, and differ in several aspects of their transposition: their reverse transcription is primed by annealing to single stranded DNA at the nicked target site, precluding the need for an integrase or VLP. Both LTR retrotransposons as well as LINEs can have non-autonomous counterparts: LARDs (LArge Retrotransposon Deletion derivatives) and

SINEs (Short Interspersed Nuclear Elements), respectively. These elements do not code for the proteins necessary for their transposition but maintain the structural characteristics and can be activated in *trans* by proteins encoded by another element. While LINEs and SINEs tend to not be very frequent in plant genomes (though have proliferated in some), they have had great success in some mammalian genomes, for example the infamous Alu SINE in the hominid lineage(Salem et al. 2003).

DNA transposons are dubbed so because they transpose via a DNA intermediate (**Figure 3**). They are flanked by short terminal inverted repeats (TIRs) and usually encode a single protein, a transposase, which recognizes the TIRs, excises the element and inserts it in its new location, creating a jagged cut over a few nucleotides which, upon repair, generates a characteristic target site duplication (TSD).

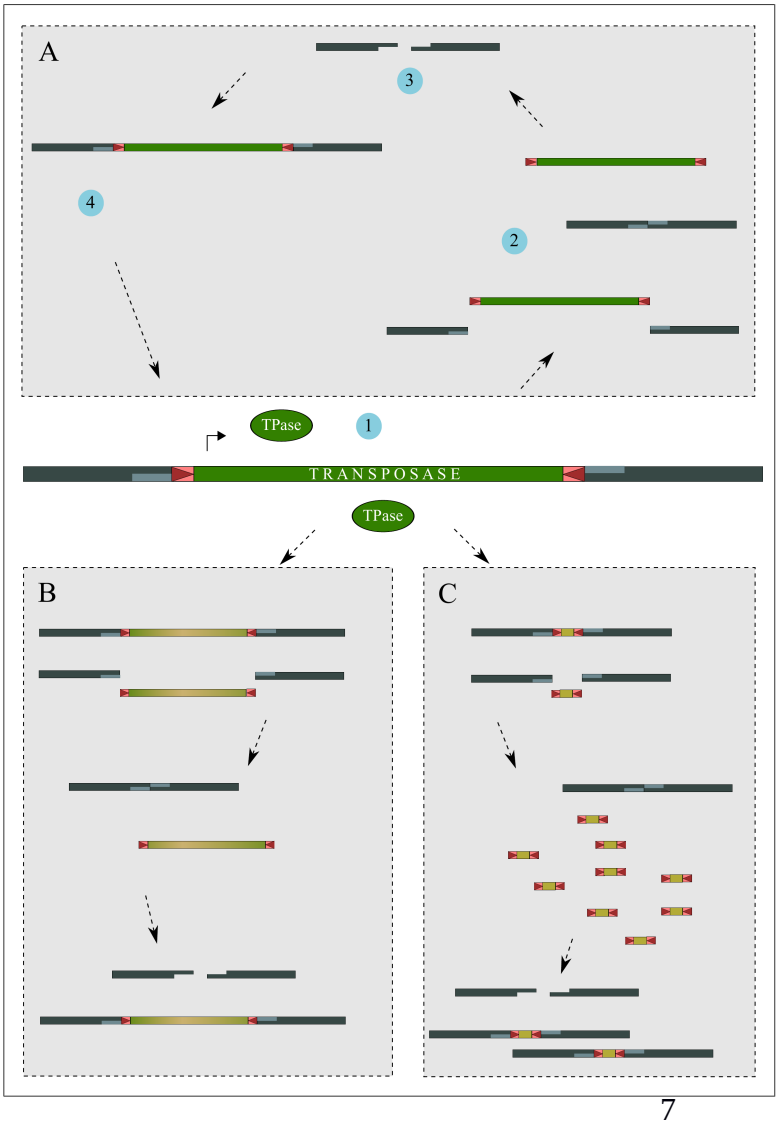


Figure 3: autonomous and non-autonomous DNA TE mobilization

A) life cycle of an autonomous DNA transposon:

1- transcription and translation of transposase enzyme

2- excision of the element and ligation of empty site

3 - jagged cut at insertion site

4 - insertion of excised element and DNA repair of jagged ends, leading to a Target Site Duplication

B) mobilization of a non-autonomous element by the same transposase

C) mobilization and amplification (by an unknown mechanism) of related MITEs

Differences in the transposase motifs, as well as the TIR sequences and the size and sequence of the TSD, allow the classification of DNA transposons into 6 main superfamilies: PIF/Harbinger, hAT, Tc1/Mariner, CACTA, MULE, and Helitron (**Table 2**).

superfamily	TIR	TSD	plants	mammals	fungi	others
Tc1-Mariner	30-225bp	TA	x	x	x	x
hAT	10-25bp	8bp	x	x	x	x
MULE	100 to 500 bp with repeats	9-11bp	x	x	x	x
PIF-Harbinger	$G(N)_5GTT$	TTA/TAA	x	x	x	x
CACTA	CACTA - - - + subterminal repeats	2-3bp	x	x	x	
Helitron	n/a	n/a	x	x	x	

Table 2: major DNA TE superfamilies

The mode of transposition of Helitrons is still not clear, and it believed to be in a rolling-circle mechanism, like some bacterial mobile elements. Nevertheless, it is classified as a DNA transposon. DNA transposons can also have non-autonomous counterparts, such as deletion derivatives of whole elements, which can be mobilized in *trans* by the transposase encoded by an autonomous element (**Figure 3 B**). The paradigmatic example is the maize dissociation element (Ds) that can be mobilized by the activator element (Ac), and that were genetically identified by McClintock long before they were molecularly characterized (Fedoroff, Wessler, and Shure 1983).

A particular type of defective DNA transposons are MITEs (Miniature Inverted-repeat Transposable Elements). These are very short (around 300-1000bp), share the TIRs of their autonomous element and sometimes some of the internal sequence as well. They are mobilized in *trans* by the related transposase and, by some unknown mechanism, can amplify and reach very high copy numbers, which differentiates them from the canonical defective elements (**Figure 3 C**).

Within each superfamily, there is a diversity of element types, and these are further classified into families on a basis of DNA sequence similarity. This final step is the most controversial, since the definition of similarity can be rather subjective. Typically one uses the criteria of 80% identity along 80% of the sequence length to place two sequences in the same family (Wicker et al. 2007). The fact

that transposon copies degenerate over time, accumulating mutations, deletions, and other rearrangements, makes a family a continuum of related sequences rather than a set of clearly similar elements. (see **Figure 1.1, Chapter 1)**).

What I would like to highlight here is the extreme diversity of transposable elements: though they share the capacity of movement, that is often the only feature two superfamilies share. While some TE superfamilies are common to many clades of evolution (such as LTR retrotransposons, for example) some are specific to a lineage (such as Alu in hominids) and many families are species-specific. This makes the question of the origin of transposable elements and their evolution a particularly interesting one.

Prevalence of TEs in genomes and methods for bioinformatic identification

The recent proliferation of genomic sequence data has generated a wealth of information for the study of mobile elements but this information has to be mined. Given the sequence of a whole genome, one has to identify which regions are related to transposons, and characterize these elements by defining their relationship in families and superfamilies in order to study their evolution. As we have seen, transposons are very diverse and this diversity makes it necessary to employ various methods for their characterization (these methods are further reviewed in Results: Chapter 1). The genome sequencing projects within the last 12 years have revealed that transposons can occupy a very large fraction of these: from 20% in the compact genome of Arabidopsis to 45% in humans to 85% in maize. This came as quite a surprise, that not only is little of the genome genes but what isn't are highly mutagenic elements. Transposons can actually be one of the major contributors to differences in genome size: for example the genomes of *Arabidopsis thaliana* and sorghum reach 120Mb and 700Mb, respectively, and the difference is mainly due to the varying abundance in LTR retrotransposons (The Arabidopsis Genome Initiative 2000; Paterson et al. 2009). Interestingly, different transposons tend to make up the major contingent in different genomes: often transposition occurs by bursts of just a few families and these reach high copy numbers. For example in barley less than a dozen LTR retro families account for over half of the genome (Wicker et al. 2009), and in rice bursts of retrotransposon activity have shaped the current state of the reference genome (Elbaidouri and Panaud 2013). Comparative

genome analyses have revealed that even between closely related species, or varieties within a given species, the transposon content can be highly variable. Bursts of transposition in a given lineage can lead to high levels of polymorphism between cultivars, such as that generated by the sudden amplification of the *mping* MITE in rice (Naito et al. 2006). The comparison of closely related varieties can yield a wealth of information regarding the activity of transposons, since it enables a picture on a very short timescale and before selective pressures have had too much time to erase evidence of transposition. These types of comparisons, across many species and lines, will hopefully yield insight into the frequency and impact of TE activity in evolution.

Impact on the host genome

Most TEs were first discovered by changes in phenotypic characters due to insertional mutagenesis. For example, the plant elements Ac/Ds corresponding to the genetic elements proposed by McClintock (Fedoroff, Wessler, and Shure 1983) or the En/Smp elements (Pereira et al. 1985), or the snapdragon Tam3 element (Hehl et al. 1991), the Drosophila elements p (O'Hare and Rubin 1983) and hobo (McGinnis, Shermoen, and Beckendorf 1983). More recently, insertional mutagenesis has continued to be the basis for discovery of new transposons, such as *mPing*, which was found inserted into the gene for *rice ubiquitin-related modifier-1* (*Rurm1*), and whose excision resulted in the reversion of the "slender glume" phenotype (Nakazaki et al. 2003) and *dTstu1*, the source of a somaclonal variation inducing purple pigment synthesis in a usually red potato variety (Momose, Abe, and Ozeki 2010). However, while insertional mutagenesis is the most obvious impact of transposition and the most easily detected by phenotype, transposons can have a range of effects on the genome they occupy and the genes they cohabit with.

On a structural scale, transposons, by virtue of being repetitive sequences, can be involved in illegitimate recombination generating chromosomal rearrangements such as translocations and inversions. Respective to chromatin organization, transposons can influence the formation of heterochromatin. In some cases, this function has been embraced and become co-opted: in *Drosophila*, which lacks telomerase, specific families of TEs form the telomeres, elongating them through

successive bouts of insertion into themselves (Biessmann et al. 1992). In addition, it has been proposed that TEs may nucleate heterochromatin in *S. pombe* centromeres, making the concentration of TEs in the centromere essential for its function (Almeida and Allshire 2005). In plants, centromere-specific TE families such as CRM in maize aid in the formation of centromeres and bind centromere-specific variants of the histone H3 (Jin et al. 2004). In a parallel manner, centromere-binding proteins such as CENP-B have convergently evolved from transposases in yeast and mammals (Casola, Hucks, and Feschotte 2008).

CENP-B is not the only example of a domesticated transposase: the examples are numerous and often it is the DNA-binding capacity and/or nuclease function that is co-opted. Recombination proteins require both DNA binding and nuclease capacities, and in humans RAG1 is believed to be derived from a now-extinct Harbinger transposon (Kapitonov and Jurka 2004). This exaptation of a transposition mechanism led to the evolutionary innovation behind the mammalian adaptive immune system, where DNA recombination generating new combinations of the V(D)J genes is strictly limited to lymphocytes (Fugmann 2010). The DNA-binding and nuclear localization capacity of transposases have also been domesticated into various transcription factors, an example of which are FHY and FHL which mediate light response in *Arabisopsis thaliana* (Hudson, Lisch, and Quail 2003). In these cases the novel protein is entirely derived from a transposase, but it can also happen that a TE inserted into a gene contributes a modular part, in a process called exonization.

Not only have the proteins TEs encode been domesticated for various purposes, but their actual DNA sequences – and variability in sequence and genomic distribution – have been sources for regulatory elements. TEs carry their own promoters and read-though or read-out transcription can induce the expression of nearby genes (Hernández-Pinzón et al. 2009). In humans, TEs have been shown to have generated and distributed TFBS for several master transcription factors (Bourque et al. 2008; Kunarso et al. 2010). The recent ENCODE project has permitted an estimation of the frequency with which TEs contribute to regulatory sequences: analysis of cell specific regulatory sites by DNase1 hypersensitivity show that they are enriched in LTR retrotransposons (Thurman et al. 2012), and 18% of the transcription start sites (TSS) overlap with repetitive elements (Khatun et al. 2012). These analyses confirm previous evidence that a large part of conserved (and therefore functional) non-coding

sequence in the mammalian lineage are TE-related (Mikkelsen et al. 2007). Whole-genome sequencing revealed to which extent TEs participate to the structure of the genome, and these results provide an insight into the degree to which these TE sequences are functional.

While TEs have incontestably contributed many functions and are a source of new variability for evolution, they remain nonetheless a very mutagenic element and as such genomes have developed various mechanisms to control them. TEs are the targets of epigenetic silencing, through DNA methylation and histone variants, as well as transcriptional and post-transcriptional gene silencing mediated by small RNAs. These mechanisms are dynamic and are regulated in various manners, and respond to various situations. Many of the silencing small RNAs come from the TEs themselves, in a sort of feedback loop. TE silencing is relieved in certain moments, for example in response to environmental stresses such as pathogen invasion (Grandbastien, Spielmann, and Caboche 1989) or heat stress (Pecinka et al. 2010) or environmental factor such as cold (Butelli et al. 2012). This has led to certain genes acquiring inducible TE-derived promoters or enhancers: some resistance genes owe their stress inducibility to TE derived promoters (Feng, Leem, and Levin 2013) as well as other vernalization-dependant expression profiles such as cold-specific expression of anthocyanins in blood oranges (Butelli et al. 2012).

Thus even the silencing mechanisms to which TEs are subjected have been co-opted for cellular functions, and it has been postulated that complex regulation mechanisms such as siRNA and miRNA were originally developed to regulate viruses and TEs, then were exapted for complex regulation of gene expression (Waterhouse, Wang, and Lough 2001; Plasterk 2002). In fact, in addition of being at the origin of this mechanism, TEs can be the source of siRNAs and miRNAs regulating gene expression (Piriyapongsa and Jordan 2007; Piriyapongsa and Jordan 2008). Additionally, TEs have also developed strategies to escape silencing and continue to proliferate (Hernández-Pinzón et al. 2012), which in some cases are based in the production by the TE of miRNAs that counteract host silencing (Nosaka et al. 2012).

The extent of our understanding of TE's contribution to genomes and evolution is broadening, and there is a plethora of examples supporting their role in evolution and generating variability. The fact

that similar functions fulfilled by exapted TEs have evolved convergently suggests that though TEs are not maintained under phenotypic selection in the short term, they might be necessary in the long term for the evolution of complex organisms, and for that reason their self-replicating mechanism of maintenance is a key aspect to their role as a well of potential diversity. What we don't know is to what extent TE activity has an influence on selection and the frequency with which it happens.

In analyzing specific examples there is a certain ascertainment bias in our vision of TE's role: we study genes that are important, and then are surprised if we see that they somehow are related to or regulated by TEs. But the question is, how often does this happen and how often is this advantageous? The analysis of closely related genomes to identify maps of polymorphisms and genome-wide characterization of exaptation events will give us a better perspective of the frequency with which these events happen. Since every genome has its own story, the more genomes we can investigate the better, and the broader our understanding of transposable elements in general and their interaction with their host.

This is the context in which I would like to place my PhD work: the goal of my dissertation has been to investigate the role of TEs in plants and their impact on gene and genome evolution. For this I have taken two approaches. The first is a study in the newly sequenced genome of *Cucumis melo*, an important crop plant in Spain. In the context of this project I have characterized the transposon landscape in the genome, and identified TE related polymorphisms between seven different varieties. This project has of interest the fact that this is an important plant for agriculture and domestication is a particularly relevant evolutionary context in which to study the impact of transposons, as the lines analyzed come from different geographic and selection backgrounds. In the context of this project I have developed a pipeline for genome annotation, and a software for detection of polymorphisms using next-generation paired-end sequencing data.

The second part of my project has been the case study of MITE families which have amplified a TF BS in the model plant *Arabidopsis thaliana*. This project focuses on the potential impact of the redistribution of this TFBS, a phenomenon that has been described for various master TF in animals but not yet to my knowledge in plants. This study has the advantages that come with working with a model plant: a very well-curated genome sequence and annotation, as well as other corroboratory data such as microarrays and extensive molecular biology evidence of gene functions.

Objectives

OBJECTIVES

The objectives of this work can be divided in three groups and are are as follows:

1. Analysis of the melon transposon landscape:

1.1 annotate transposons in the genome of melon (*Cucumis melo*), using available and developing novel bioinformatic tools.

The goal of this study is to discover the transposon landscape in the melon species, and gain insight, by studying the families present and characteristics of their copies, into the history of transposition that has led to this state.

1.2 Study the melon transposon content and compare it with that of the related species cucumber (*Cucumis sativus*).

Analysis of the closely related species cucumber, and comparison of the specific and shared TEs in these two genomes, will yield information as to how these have evolved in each genome since the divergence of the two species.

2. Analysis of the contribution of transposons to the recent evolution of melon.

2.1. Develop bioinformatic tools to study transposon movement using next generation sequencing data.

Next generation sequencing of varieties offers the opportunity to perform comparative genomics for structural variation detection, and I have set as a goal to develop a tool to specifically detect TE-related polymorphisms.

2.2. Analyze transposon insertion/deletion events in melon varieties.

The objective is to use this tool to construct a map of polymorphic sites in seven melon lines, and analyze their potential impact on gene and genome evolution, completing the picture drawn from the annotation of the reference as to recently mobile elements.

3. Analysis of transcription factor binding site amplification by transposons.

3.1. Study the possible amplification of the E2F binding site by MITEs in *Arabidopsis thaliana*.

The goal of this project is to describe the capture and amplification of a master TFBS in Arabidopsis, and assess the impact this may have had on gene expression.

Materials and Methods

MATERIALS AND METHODS

Materials and Methods for Chapter 1

Dating insertion time of LTR retrotransposons

For this analysis we considered only the families that have more than 10 copies that cover at least 90% of the length of the family representative. For each of these families, we aligned these long (>90% query coverage) elements to the representative and selected those which aligned with at least 50% of the length of the representative's LTRs, as defined by LTR_FINDER. The two LTRs of each selected element were aligned and the date of divergence calculated using Kimura's two-parameter method (Kimura 1980):

let P be the transition fraction in the aligned sequences
Q be the transversion fraction
K be the the evolutionary distance
T be the time of divergence
k be the evolutionary rate

then $K = -1/2 * \ln[(1-2P-Q) * \sqrt{(1-2Q)}]$

and $T = K / 2k$

We took k as 1.3 e-8 substitutions/site/year which has been used to date LTR retrotransposons in this manner (Choulet et al 2010) and is taken from the rate calculated for the *Adh* locus in grasses (Gaut et al 1996) and adjusted (divided by two) for the fact that LTR retrotransposons display a higher substitution rate than genes.

Phylogenetic analysis of melon TE families

The protein-coding regions of all copies covering at least 50% of their respective query were extracted using tblastn with a protein query corresponding to the superfamily's transposase. These sequences were then aligned with ClustalW (http://www.clustal.org/clustal2/) and phylogeny reconstructed with phyML (http://www.atgc-montpellier.fr/phyml/)

Material and Methods for Chapter 2

Sequencing data of melon varieties

Libraries of Illumina paired-end reads (500bp fragment length, 150bp read length) for seven cultivars and the reference DHL92 were obtained from the following sources:

Cultivar	Reads	Reference
DHL92	35,538,240	Garcia-Mas et al 2012
PS	35,857,911	Garcia-Mas et al 2012
SC	35,233,293	Garcia-Mas et al 2012
CV	28,038,962	Gonzalez et al 2013
IRK	33,207,205	Gonzalez et al 2013
VED		
TRI		
CAL		

DHL92 is a doubled haploid line derived from PI 161375 x T111 and represents the melon reference genome.

Quality filtering of paired-end Illumina reads

Quality of reads was assessed with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) then filtered with SGA (https://github.com/jts/sga) requiring a base quality of at least 20.

Paired end read mapping

Filtered reads were mapped to the assembled reference (CM3.5, https://melonomics.net/) using BWA(Li and Durbin 2009) (http://bio-bwa.sourceforge.net/) with aln parameters: -n 6 -o 1 -e 1, sampe default parameters.

Detecting deletions in the resequenced sample with respect to the reference

Pindel (https://trac.nbic.nl/pindel/wiki/WikiStart) with default parameters was used to detect deletions in the melon lines with respect to the reference. Predictions with less than 2 supporting forward and reverse reads were discarded, as well as those predicting a deletion less than 200bp.

jitterbug installation and usage

Algorithm description

The input to the program is a .bam file of the sample's reads mapped to the reference genome, and an annotation of TEs in the reference. The algorithm is implemented in python, and follows four main steps:

0. Calculate mean and standard deviation of insert size (fragment length) over 1 000 000 properly paired read pairs

1. Select the discordant reads from the .bam file. For this, scan the bam file and reject any read pair that is flagged as “proper pair” (sam bitwise tag 0x2), or that has a mapping distance less than the expected insert size, or where both reads in a pair are mapped repetitively. The reads that are left are written to a bam file of “valid_discordant_pairs”. All softclipped reads are saved to a separate file, to be used in future validation steps. [** check this!!] This step uses the pysam module.

2. Of the valid discordant pairs, select those that have one read mapping uniquely to a non-TE location (“anchor” read), and the other read mapping (repetitively or not) to at least one location that is annotated as a TE in the provided annotation (“mate” read). This step uses the pybedtools module. The selected reads are returned as a list of AlignedReadPair objects. For each AlignedReadPair, the interval of putative insertion site corresponding to this read pair is of length the fragment length (+ xSD, x set by parameter -d) and in the direction to which points the read that maps uniquely to a non-TE location.

3. The AlignedReadPair objects are clustered into maximal clusters according to the overlap of their predicted insertion interval, on the forward and the reverse strand. This is implemented in the ClusterList class, using the Cluster object. The reads that compose these clusters are written to the “final_clustered_reads” bam file
Finally, clusters are paired as one forward and one reverse cluster into a ClusterPair object, if their predicted insertion intervals overlap. This step is parallelized by chromosome, where the number of threads used is set by the -p option. The properly paired reads which fall in the predicted interval are scanned for softclipped positions and then for reads that overlap that position in order to determine the zygosity.

4. Each cluster pair calls one putative TE insertion, with the insertion site falling within the intersection of the forward and reverse predicted intervals. These are the insertions written to the “TE_insertions_paired_clusters.gff” file
If there are unpaired clusters, these are written to the “TE_insertions_single_cluster.gff” file.
Tables with more information about the reads that compose these clusters are written to the respective “supporting_clusters.table” files.

Usage

requirements

Jitterbug requires python version > 2.7.3 (has not been tested with python 3) and the following modules:
pysam (version 0.7)
pybedtools (version 0.6.1)
numpy (standard with python version 2.7.3)

The "helper" classes (AlignedReadPair.py, ClusterList.py, BamReader.py, Cluster.py, ClusterPair.py) should be in the same directory as the main script (Run_TE_ID_reseq.py), or in the PYTHONPATH

installation

you can clone the sourceforge repository by:

```
git clone ssh://username@git.code.sf.net/p/jitterbug/code jitterbug-code
```

usage

```
TE_ID_reseq -i bam_file.bam -t TE_annotation.gff -l lib_name
```

options:

-i bam_file of aligned reads, ordered by pairs. This is the default order for alignments generated by bwa. If this is not the case, you can sort it by name using

```
samtools sort -n bam_file bam_file.nsorted
```

-s [True|False] if set to True will only consider reads that map repetitively to a TE in step 2. default False. WARNING: whether a read that maps to several locations in the genome is actually flagged as such in the bam file depends on the mapping parameters you used. (in bwa, the sampe -n and -N, for example). Only set this option if you are sure that your mapper is not throwing away the alternative mapping information.

-v [True|False] if set to True will print a (very) wordy output about what is being done

-t Annotation in gff3 format of the transposons in the reference

-c (int) min cluster size to be considered. Default 2.

-l (string) library name, to be used in final gff output

-d (int) multiplicative to be used when calculating the insertion intervals for read pairs in step 2. if -d is set to x, will calculate the the insertion interval as xSD + fragment_length.

Default 2

-o (string) prefix for output files. If not set, will use name of input bam file

-a [True|False] use an already filtered bam file, in order to skip step 1. If set, will look for a file named <prefix>.valid_discordant_pairs.bam. <prefix> is either the input bam file name, or the prefix given in -o . default False

-n (string) name of the tag present in the gff of TEs to use to record the TE annotations that are identified as inserted sequences. Default Name

-p (int) if set, will parallelize the cluster calculation using the specified number of threads. Currently this only works with multicore computers, not a cluster, and requires the pp (parallelpython.com) module.

Output

bam files

Three bam files are output, which are all subsets of the original input bam file. They are:

<prefix>.proper_pair.bam

which are all the properly paired reads in the original bam file. These are used to look for the softclipped reads that may indicate the exact insertion site within the predicted interval.

<prefix>.valid_discordant_pairs.bam

which are all the reads selected in step 2, and that were considered for clustering

<prefix>.final_clustered_reads.bam

which are all the reads that were used in forming the final cluster pairs.

The last two bam files can be useful for visualizing the reads upon which the predictions have been based.

gff annotations of called TE insertion sites

The called TE insertions that correspond to a pair of clusters (one forward, one reverse) which overlap in their prediction interval are output in gff format to the file

<prefix>.TE_insertions_paired_clusters.gff3

The tags in the 9th column are:

Inserted_TE_superfam_[fwd|rev]

if the provided TE annotation had a tag named superfam, this tag will record that value of all the TE annotations that the fwd and rev clusters point to, respectively. Otherwise, undefined.

Inserted_TE_names[fwd|rev]

same as previous, except that its the value of the tag specified in the -n parameter

supporting_[fwd|rev]_reads
number of reads that constitute the fwd and rev clusters, respectively

cluster_pair_ID
unique identifier

lib
library name, taken from -l parameter

[fwd|rev]_cluster_span
range spanned by the start positions of reads in fwd and rev clusters, respectively. A span of 0 means the reads are stacked

softclipped_pos
tuple: (start, stop) of the positions within which one or more reads were softclipped. This interval is putatively the exact (+/- 3bp) position of insertion.

softclipped_support
number of spftclipped reads supporting the previously mentioned position

het_core_reads
number of properly mapped reads that span the interval specified by softclipped_pos. These reads would indicate the presence of a non-insertion allele.

zygosity
ratio of softclipped_support / het_core_reads. A value around 0.5 would indicate heterozygosity, a value near 1 nomozygosity for the insertion.

Additionally, a gff file of unpaired (single) clusters is written to:
<prefix>.TE_insertions_single_clusters.gff3
though these are highly unreliable predictions

tables of cluster descriptions

Additional information regarding the TE insertion predictions is output in table format as:
<prefix>.TE_insertions_paired_clusters.supporting_clusters.table

The format of the table is described in its header:

this table describes the read clusters identified in the bam file input.bam and corresponding to the transposon annotations in TE_annot.gff

parameters: -s False -c 2

this table contains three types of lines, tab-delimited:

- insertion lines: one per predicted insertion site, corresponding to a pair of overlapping clusters, one fwd, one rev

columns:

I	indicates this line describes an insertion interval
cluster_pair_ID	unique ID of this prediction, common to the R and C lines of this cluster
lib	library name, from the -l parameter
chrom	chromosome
start	start position of prediction interval
end	end position of prediction interval
num_fwd_reads	number of reads in forward cluster
num_rev_reads	number of reads in reverse cluster
fwd_span	span of forward cluster
rev_span	span of reverse cluster
best_sc_pos_st	start position of softclipped interval
best_sc_pos_end	end position of softclipped interval
sc_pos_support	number of softclipped reads supporting this interval
het_core_reads	number of reads mapping over softclipped interval
zygosity	ratio of het_core_reads / sc_pos_support

- cluster lines (two per insertion, one fwd and one rev):

columns:

C	indicates this line describes a cluster
cluster_pair_ID	unique ID of this prediction, common to the R and I lines of this cluster
lib	library name, from the -l parameter
direction	[fwd|rev] indicates this cluster is composed of reads mapped on the forward or reverse strand
start	start position of cluster
end	end position of cluster
chrom	chromosome
num_reads	number of reads composing the cluster
span	span of the cluster

span is defined as the range of start positions in the cluster. A span of 0 means that all the reads originate at the same start site, and are probably an artifact. A span the size of the fragment length indicates good coverage.

- read lines (fwd reads constitute the fwd clusters, rev reads the rev clusters)

the reads that are "anchor" are those that constitute the cluster, the reads that are "mate" are the anchors' mates, which map to a TE

columns:	
R	indicates this line describes a cluster
cluster_pair_ID	unique ID of this prediction, common to the C and I lines of this cluster
lib	library name, from the -l parameter
direction	[fwd|rev] for anchor reads, indicates whether it is mapped on the forward or reverse strand. For mate reads, is the opposite of its corresponding anchor
start	start position of read mapping
end	end position of read mapping
chrom	chromosome
type	[anchor|mate]
bam_line	columns corresponding to the original line in the mapping bam file.

Materials and Methods for Chapter 3

Annotation of TEs in Brassicae

The annotation of TEs in *C. rubella*, *T. halophila*, and *B. rapa* was performed with RepeatMasker (http://www.repeatmasker.org/) using the *Arabidopsis thaliana* repeat database dowloaded from RepBase (www.girinst.org). MITEs were annotated with SUBOTIR (Jordi Payet, unpublished) and the predictions of both were merged to obtain a non-redundant annotation. We chose this method since these genomes are very close to Arabisopsis and there is an extensive repeat database available for this genome. We did not annotate LTR retros or do any other kind of *denovo* annotation because we know that the elements we were looking for are DNA TEs.
The TE annotation for *Arabidopsis lyrata* was taken from Hu et al. 2011 and that of *Arabidopsis thaliana* from www.arabidopsis.org, version TAIR9.
The families SimpleHat1, SimpleHat2 and SimpleGuy1 were re-annotated in *A. thaliana* by:
- aligning the elements of a given family as defined by TAIR9
- taking the borders, conserved regions without the minisatellite in between and concatenating them, using this as a query for copy-finder, allowing up to 10000bp gap.
This allowed to join some fragmented copies in TAIR into full MITEs with the minisatellite in between, and identify some copies with fragments of TPase

Identification of E2F binding motifs

The coordinates of the E2F sequence TTCCCGCCAA were identified with vmatch (http://www.vmatch.de/) for perfect matches on either strand, or with edit distance of 2 when identifying positions of E2F-like sequences.

Identification of minisatellites

Tandem repeats were identified with TRF (http://tandem.bu.edu/trf/trf.html) (Benson 1999) with the following parameters:

2 5 5 80 10 30 35

These differ from the default parameters in the:
- second and third parameters: mismatch and indel penalty, respectively (decreased)
- sixth parameter: min alignment score to report (decreased)
- last parameter: max length motif to report (decreased)

These parameters were modified from default in order to identify more degenerate tandem repeats (more permissive alignments and minimum score) and in order to only report tandem repeats with shorter periods (since we know the minisatellite we are looking for has a period of approximately 27bp)

Annotation manipulations

Intersections and overlaps of sets of annotations were performed with the BedTools suite (http://code.google.com/p/bedtools/) (Quinlan and Hall 2010)

ABBREVIATIONS

TE: transposable element
RT: reverse transcriptase
LTR: long terminal repeat
TE: transposable element
TP: transposase
SO: Sequence Ontology
TIR: terminal inverted repeat
HSP: High Scoring Pair
ORF: open reading frame
HMM: Hidden Markov Model
MITE: Miniature Inverted-repeat Transposable Element

Results

Chapter 1

CHAPTER 1: TE ANNOTATION AND ANALYSIS IN THE MELON GENOME

1.1 Introduction: complexity of transposon annotation

1.1.1 Background

Transposons were first identified due to to their impact on phenotype and the extent to which they are prevalent in genomes remained difficult to grasp until the advent of whole-genome sequencing, now our vision of transposons is broadening from just a few elements in select species to genome-wide annotations of the many available genomic sequences. This has revealed that transposons occupy quite a large fraction of both plant and mammalian genomes, from 20% in the model plant *Arabidopsis* (www.tair.org), to 45% in humans (Venter et al. 2001) and 85% in maize (Schnable et al. 2009). These data show the diversity transposon types as well as prevalence, as the families present in different genomes can differ greatly, and though the same type might be identified in different genomes they can have amplified to varying extent. However for genomic sequence data to be useful for the study of transposons, one needs to know which regions of the sequence are related to transposable elements, as well as their family and superfamily classifications. In this consists the task of genome annotation.

Annotation of transposons is not straightforward, as they are varied both in structure and function, and their diversity makes even their classification an issue of debate. A nomenclature system has been suggested by Wicker et al (2007) that separates transposons into classes, according to their transposition mechanism (Class I via an RNA intermediate, and Class II for a DNA intermediate) then further into superfamilies according to broad features (such as the structure of encoded proteins or non-coding regions, or target site duplication (TSD) length) then finally into families according to sequence similarity . Even the definition of a family can be onerous since individual elements are subjected to point mutations, deletions and insertions that make a TE family a continuum of related sequences and fragments rather that a group of clearly similar sequences. Typically one uses the criteria of 80% identity along 80% of the sequence length to place two sequences in the same family. (Wicker et al.

2007)

A transposon might have a well-defined structure, or encode a protein, or both, or neither. In some cases the structure is easily identifiable in genomic sequences, such as LTRs which are generally long (greater than 500bp). Other structures are just as well defined but are too short to be informative, such as TIRs of many Class II transposons, or can be specific to a given family such as the 3' region of SINEs. In some cases an encoded protein is conserved over several superfamilies, such as the reverse transcriptase (RT) of LTR retrotransposons, in others it can be variable such as the transposase of Class I elements (Wicker et al. 2007) However, one characteristic that is shared by most (though not all!) is to be found in multiple copies in the genome.

For these reasons there is no standard method for identifying all transposons in a genome in one step. The approach used first depends on the objectives of the study. Indeed, transposons by their repetitive nature and coding capacity pose problems to functional analyses like gene prediction and need to be masked prior to running these predictions. Thus some methods of transposon annotation are designed to mask these sequences, not necessarily to annotate them in a manner that enables their analysis. It is indicative that one of the first tools for transposon analysis is called RepeatMasker (http://www.repeatmasker.org). However, if one wants to annotate transposons to study them, the annotation has to be performed in a way that provides information on the families of elements as well as annotating the structural characteristics of each element as completely as possible.

Model species such as human, *Arabidopsis* or *Drosophila* have had their genome fully sequenced for a long time, and the transposon annotation has gone through several revisions and manual curation. However, when one is presented with a new genome, especially one that does not have a close relative sequenced and well annotated, one is faced with the choice of different strategies to annotate TEs.

There are a few aspects of TE biology that must be kept in mind when developing or evaluating computational methods for TE annotation. Indeed, individual TE instances may be fragmented due to recombination between elements, nested insertions, or incomplete retrotranscription, among others, and

thus may only present partial similarity to other members of its evolutionary family. Also, sequence divergence causes individual elements to evolve separately and resolving families is quite dependent on the thresholds of similarity used to define families. These factors make the definition of families very dependent of the algorithmic methods and similarity thresholds used.

The task of TE annotation consists in two types of analysis: TE discovery and TE identification. The former aims to discover TE elements in a given sequence, the latter to identify sequences related to a given TE. Both are necessary for a full annotation of the transposon landscape in a genome.

1.1.2 Methods for TE annotation

De novo annotation based clustering repetitive sequences

One approach, which bundles both discovery and identification, is to exploit the fact that transposons tend to be present in large copy numbers, and scan the genome for repeated sequences without using any prior information in regards to TE structure or similarity to known TE sequences. This has the advantage of potentially identifying transposons unique to this genome, but also several challenges. First, there is the pitfall of mis-annotating other types of repeats as transposons. Indeed, there are many repetitive sequences throughout the genome that are not transposons, for example centromeric repeats, tandem repeats or segmental duplications. Second, TE families composed of largely non-overlapping fragments or present in low copy number will be overlooked by these methods. The final challenge is the classification into families of the sequences thus identified, due to the aforementioned diversity of elements within a family which makes clustering these sequences difficult. This strategy has been implemented by software such as RepeatScout (Price, Jones, and Pevzner 2005), but that has been shown to be rather unspecific when benchmarked against the curated annotation of the *A. thaliana* genome and recover only fragments of the elements it correctly identifies (Flutre et al. 2011).

TE identification based on representative elements

Another approach is to first discover TEs in your given genome (for which many methods are possible), then identify all the individual copies comprising the families of these elements. In this case, the discovery step aims to find "representative" elements, which would be the minimal set of sequences that represent the diversity of elements in the genome. A representative sequence would thus be one which, when used as a query to identify similar sequences, could retrieve the maximal number of fragments in its family (**Figure 1.1)**

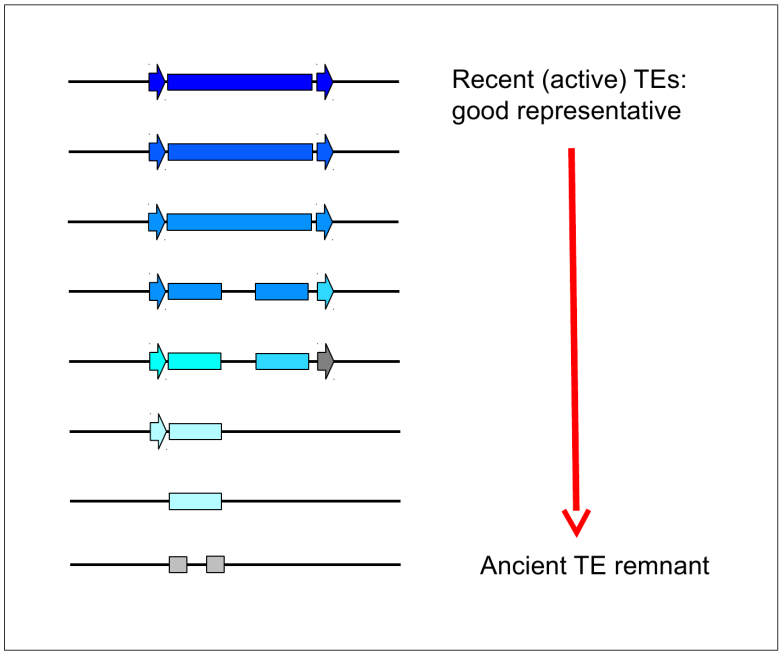


Figure 1.1: A TE family is a continuum of similar sequences

RepeatMasker, a widely used tool to identify transposable element sequences, bypasses the discovery step and uses as representatives consensuses of elements found in other genomes (usually taken from the RepBase database (http://www.girinst.org/repbase/)). While this approach might be sufficient for masking the most conserved regions of TEs before running gene predictors, is has been shown to be "neither the most efficient nor the most sensitive approach" for TE annotation (Juretic, Bureau, and Bruskiewich 2004), and that representative sequences that are specific to a given genome are more apt as queries to recover their given family members than consensus sequences, as these tend to not include

the specific non-coding sequences and structural characteristics (Buisine, Quesneville, and Colot 2008).

There are several approaches to identify representative TE sequences in a genome, either de novo, homology-based or structure-based.

Methods for representative discovery

De novo methods are based on the identification of repeated sequences (for example by whole-genome self-alignment) then clustering and categorization. This approach has been evaluated by Flutre et al. (2011) by benchmarking against the *A. thaliana* and *D. melanogaster* annotations. These authors compare the performance of different algorithms for whole-genome self alignments (BLASTER and PALS) and clustering (GROUPER, RECON and PILER). They then classify the elements based on structural characteristics and/or coding capacity, discarding sequences that do not display any TE features as false positives. This final step is essential for increasing the specificity of the annotations, and is what distinguishes their work from previous implementations of this approach, however it also potentially eliminates any completely novel TE families that would have different characteristics from any known TE. They also show that representatives thus identified perform just as well but not better than finding copies of well-curated representatives.

Homology-based methods use the knowledge base of the large number of TE sequences that have already been characterized, and the fact that coding sequences tend to be well conserved over certain types of elements. Indeed, the RT proteins of LTR retrotransposons are generally conserved as are certain domains of TPase of DNA transposons (Wicker et al. 2007). Transposon related sequences are available in general databases such as NCBI (http://www.ncbi.nlm.nih.gov) which one can retrieve with key terms such as “transposase” or “retrotransposase” and there are also transposon-specific databases such as RepBase for all types of transposons, or GyDB (http://gydb.org) for retroelements, or RetrOryza (http://retroryza.fr/) for retroelements in rice. Similarity search is usually implemented by local alignment search algorithms such as BLAST using protein queries against genomic sequences or

HMMs constructed from multiple alignements (Juretic, Bureau, and Bruskiewich 2004). The possibility of using HMMs is dependent on having sufficient TE sequences already characterized in the genome of interest in order to construct profiles based on alignments, so while HMM based search tools can be more sensitive than alignment based tools (Juretic, Bureau, and Bruskiewich 2004) they are not feasible in all cases. The homology-based strategy has the advantage of generating few false positives, and being capable of retrieving single-copy elements. However the drawbacks are that only the well-conserved regions of a given element will be identified and older, more degenerate elements or copies that have no coding capacity (such as MITEs or SINEs) will be overlooked. In order to characterize the full sequence of an element thus retrieved one must use other methods to identify the non-coding or less conserved regions surrounding the coding region. This can be done either by aligning multiple genomic hits, along with their flanking sequences, and by defining the borders as where the alignment breaks down, or by searching for structural elements such as TIRs in the flanking regions.

Another approach to discovering TEs in genomic sequences is exploiting characteristics specific to a given type or superfamily, and are as numerous and varied as the types of TEs themselves (reviewed in Bergman and Quesneville 2007). These methods are based on identifying a structural characteristic of a TE sequence, such as the long terminal repeats in LTR retrotransposons or the short inverted repeats of MITEs. Multiple tools implement a search for LTR retrotransposons based on identifying direct repeats within a certain window. LTR_FINDER (Xu and Wang 2007) is the most recent of these and has the advantage of allowing user-specified thresholds of divergence between the two LTR sequences as well as identification of ORFs in between them, which aids at filtering out false positives. MITEs also lend themselves to identification by structural characteristics as they are short sequences flanked by direct repeats, and found in large copy numbers. Methods for identifying these are reviewed in (Guermonprez et al. 2013) and the most recent is MITE-hunter (Han and Wessler 2010*)*. This software is the most sophisticated in that it provides several methods of eliminating false positives, at various steps of the algorithm. Similarly to others (MUST (Chen et al. 2009*))*, the first step is to identify candidate MITEs based on TIRs and TSDs. In a subsequent step candidates are discriminated based on copy number by pairwise comparison – elements that do not align with any other are eliminated as false positives. Then a consensus sequence is generated for each family and the definition of its borders verified by multiple sequence alignment with its copies taken with flanking regions. This

last step relies on the fact that within a certain family, the copies' terminal sequences (i.e. TIRs and TSDs) will be near identical and align well but the alignment will break down at the flanking regions as each element is inserted in a different genomic context. (See **Figure 1.3**)

The key to using these structure-based methods is implementing good filtering strategies to eliminate false positives, as these types of structures can occur by chance in the genome with more or less high frequency. For this reason autonomous DNA transposons are not usually discovered with this approach, even though they also have TIRs just like MITEs. Indeed, they are more easily retrieved with homology-based methods and then the search for structural characteristics is limited to their flanking regions.

Methods for identifying copies of a representative

Once one has identified representative TEs using any of the aforementioned methods, the next step is to identify the copies in their respective families. This step is necessary since a family can be composed by fragments and degenerate elements that would not have been identified by the previous step. The way this is implemented depends on your goal: if you desire only to mask the genome, it is sufficient to do a simple similarity search with a program such as BLAST or FASTA, or HMM, to identify sequences similar to your queries. However, if your goal is to study the biology of TEs, you must take into account some of the biological factors of TE evolution to get useful data. The main problem with finding copies of an element is that similarity searches will give fragmented annotations if the target sequence has a large insertion or deletion or has diverged sufficiently. Therefore in order to have a proper annotation one has to chain fragments together to properly define a copy. This is implemented in the context of certain pipelines for example MATCHER in the REPET pipeline (Quesneville et al. 2005; Flutre et al. 2011) which uses a dynamic programming algorithm to chain collinear fragments and then resolve overlaps. Another program is Greedier (Li, Kahveci, and Settles 2008) which uses a graph-based algorithm to link fragments based on maximizing its total alignment score. Both of these algorithms are embedded in their respective pipelines and cannot be used independently.

1.1.3 Objectives

It is important to note that all these analyses based on sequence comparison – similarity searches, clustering, sequence alignments – are parametrized and that the parameters chosen can greatly affect the output. For example, more lax alignment parameters can change how many clusters or families are defined, or whether a given sequence stretch is identified as a degenerate fragment of a TE or discarded. Thus, the parameters one uses must be chosen in light of the purpose of the annotation – stricter parameters for annotation in view of studying TEs, more lax if the aim is to mask as many as transposon-related sequences possible. Our goal for the annotation of the melon genome was to perform a careful and accurate annotation of the most abundant types of transposons found in plants: LTR retrotransposons, MITEs and the major families of DNA transposons. Since the objective is to study the evolution of these elements and their family relationships, we preferred to use stringent criteria in our annotation, knowing we will lose a fraction of the more degenerate elements.

1.2 Pipeline developed for annotation of transposable elements in genomic sequences

1.2.1 General strategy for annotation

The strategy followed for annotation was to first identify representative TE sequences with different methods adapted to the type of element, and then to find their copies genome-wide. We chose these methods over the *de novo* characterization of repeated sequences in the genome because we are more interested in having an accurate annotation of TEs and their characteristics than in masking all possible transposon-related sequences. We decided to focus on LTR retrotransposons, MITEs and the major families of DNA transposons using dedicated methods to identify each of these types and define as well as possible their structural characteristics. The methods for finding copies allow us to identify deletion derivatives of these representatives, thus LARDs, TRIMs and MITEs can be identified indirectly in this manner. LINEs and Helitrons were searched for using RT and helicase protein queries, allowing us to identify their most conserved coding regions.

Detection of LTR retrotransposons representatives

Since LTR retrotransposons are flanked by long (usually of more than 500bp) direct repeats, these structures can be easily identified by scanning the genome for direct repeats flanking a relatively short (few Kb) sequence. For this I used the software LTR_FINDER (Xu and Wang 2007). False positives may occur by chance, and these can be eliminated by verifying that a given putative element is present in multiple copies. Here I used a minimum of 2 copies for a sequence to be considered as a putative LTR retrotransposons. Autonomous retrotransposons encode a number of proteins needed to complete their life-cycle and therefore the lack of coding capacity for a protein similar to those (such as RT) was also used to discard false positives. This criteria would prevent us from detecting non-autonomous retrotransposons such as LARDs and TRIMs, but these can be recovered as deletion derivatives of complete elements in the copy-finding step. However, we realize that families that do not contain at least one full-length element might be overlooked, or at least under-annotated as mere fragments.

Detection of DNA transposon representatives

To identify DNA transposon representatives it is difficult to rely on their structural characteristics as their terminal inverted repeats (TIRs) tend to be short and thus less infomative for *de novo* searches which yield a high proportion of false positives. For this reason I chose a homology-based method to retrieve the most conserved TPase-coding sequences combined with a multiple alignment step in order to extend the borders of the elements. This was done by clustering similar elements and aligning them taken with flanking sequences, exploiting the fact that similar elements will align until the genomic context is reached, which is different for each insertion site (**Figure 1.3**). The advantage of this method is its high specificity but the disadvantage is that it will only identify elements found in at least two copies, as this is necessary for the alignment step.

Detection of MITE representatives

The strategy to identify MITEs is necessarily different than that used for DNA TEs since, while they also have TIRs like DNA transposons, the elements are much shorter and do not include any coding sequence. For this reason they are more difficult to detect by homology since their sequence

varies between families and genomes. They have the particularity of being found in large families, for example the *Tourist* element is found in 33,000 copies in the rice genome (Naito et al. 2006). This feature can be exploited for discovery by bioinformatic methods, by scanning the genome for short inverted repeats found close together then keeping only the sequences that are found in large copy numbers. We used a combination of two softwares, MITE_HUNTER (Han and Wessler 2010) and SUBOTIR (Jordi Payet, unpublished) to identify MITEs in the melon genome.

Identification of copies

Using these methods I constructed a database of melon transposons to use as representative sequences with which to identify families of mobile elements in the genome. I then identified all the copies, both full length and deletion derivatives, of these representatives. For this I developed a graph-based algorithm (COPILIST-NR) to join fragmented hits into maximal copies of the query sequence and resolve possible redundancies when the same genomic location is identified as a copy of distinct representatives. This approach gives us a good picture of the structures of families present in this genome as we maintain information of the relationship of each copy to the representative.

Our goal was to perform a careful annotation which would permit the use of these data for further analyses regarding the nature and evolution of transposable elements in this genome. For this reason we chose methods for transposon discovery aimed at specificity rather than sensitivity, and strict criteria for eliminating false positives, accepting the cost of losing some true predictions as well. As mentioned before the results of analyses based on sequence comparison can vary according to the parameters and cutoffs chosen, and we were very strict in both as can be seen in the methods described below.

The following flow chart in **Figure 1.2** illustrates the pipeline used.

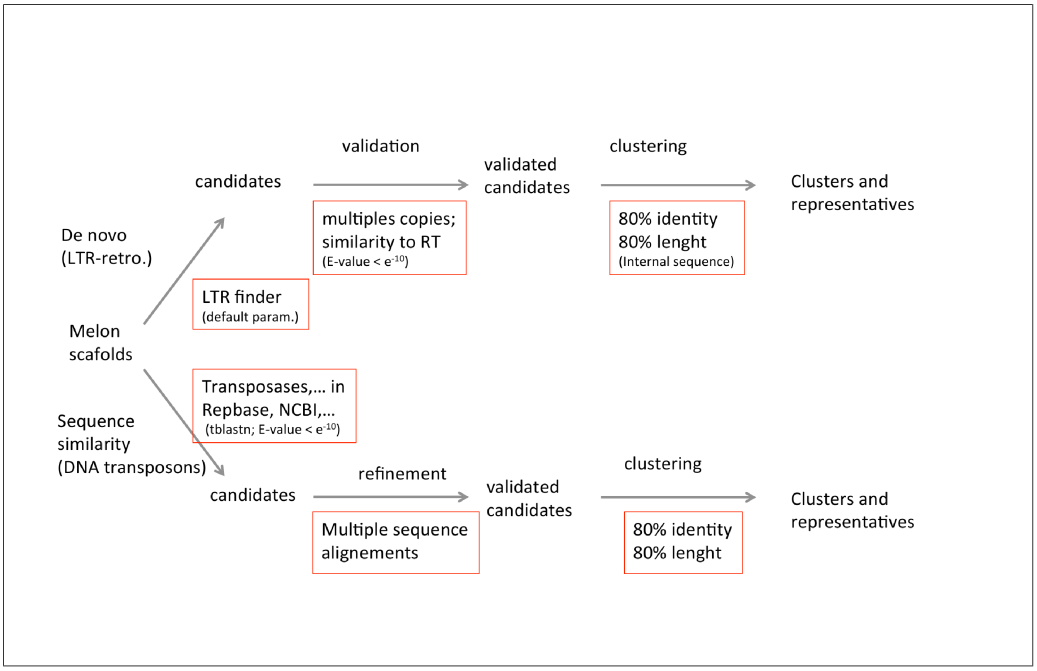


Figure 1.2: pipeline used for the annotation of transposable elements in the melon genome

1.2.2 Representative identification

LTR-retrotransposons

Candidates for LTR retrotransposons were identified using LTR_FINDER (Xu and Wang 2007) with default settings. Copies of these candidates were retrieved with a modified version of a script taken from the MITE_HUNTER suite (Han and Wessler 2010) which automates a blast search and extracts the sequence hits in fasta format. Each candidate that retrieved at least one copy was aligned with its copies using MUSCLE (Edgar 2004) (which was used for all other alignments mentioned), taking 60 bp of flanking sequence. In order to define the limits of the elements, these alignments were checked for target site duplications and that the borders of the elements align while the flanking sequences do not (**Figure 1.3 part 2)** To further verify these candidates, they were used to query a database of all LTR-retrotransposons in RepBase (www.girinst.org) with tblastx (e-value < e-10, blastall

suite available at www.ncbi.nlm.nih.gov; all subsequent BLAST analyses are performed with this suite as well). According to the best hit the candidates were attributed to either the *gypsy* or *copia* superfamilies, or discarded if no homology was found.

These verified candidates were clustered according to the internal sequence (between the LTRs as defined by LTR_FINDER), with a threshold of 80% similarity along 80% length. For this I implemented a hierarchical clustering algorithm in python and considered only columns without gaps in the calculation of percent similarity. The longest sequence of each cluster was chosen as representative sequence. These representative sequences were deposited in a database available on the website of the melon sequence project (https://melonomics.net/), with annotated features as defined by LTR_FINDER, superfamily, and copy number of the family.

Non-LTR retrotransposons

Non-LTR retrotransposons do not have the structural characteristics like LTRs which enable their *de novo* identification and therefore cannot be identified with the same approach. I had previously identified a few of them in melon BAC sequences by homology to RT (Gonzalez et al. 2010) and used these as queries to search for copies as described below. These are all annotated as "non_LTR_retrotransposon" and also link to the representative sequence they were identified with. We decided not to further describe these in the whole genome because our analysis of the BAC sequences indicated there were very few of them, as is the general case in plants.

DNA transposons

In order to identify DNA transposons, the general strategy was to fish for sequences homologous to a known transposon coding sequence and further, extend the TE sequence by aligning similar hits taken with flanking sequence, then select representatives and search for copies.

I constructed a protein database by querying NCBI with the keyword "transposase" in conjunction with transposase superfamily names such as "PIF", "hAT", "CACTA", "MULE", "hop",

“jittery”, “Mariner”, as well as “helitron helicase” to retrieve transposase sequences that have been attributed to a superfamily as well as those that have not. I excluded from these searches any sequence annotated as “putative” or “hypothetical”, in order to minimize the propagation of errors or uncertainties common in the public databases (in particular for TE sequences). I then retrieved all sequences in the genome that are similar (tblastn, e-value < e-10) to any in the transposase database. These sequences were then re-blasted against the subset of the database that have been attributed to a superfamily, and grouped according to this criterion.

The sequences in each superfamily group were clustered using UCLUST (Edgar 2010). The thresholds used for clustering were 80% similarity over 80% of the query length, which is what has been suggested by Wicker et al. (2007) as the definition of a family when comparing DNA sequences. The percentage similarity was calculated by counting any mismatch or gap as a difference, except for terminal gaps (UCLUST 'iddef' parameter set to 2) which is appropriate when aligning sequences of different length, as is the case for aligning fragments against a longer element. For selected clusters (most homogeneous or largest) the sequences were extended 5000 base pairs in either direction, and aligned. These alignments were manually inspected to extend the definition of the elements as far as the alignment was maintained, allowing to identify TIRs in some cases (**Fig 1.3**) .

Figure 1.3: Discovery of DNA TEs by homology to TPase and refinement by multiple alignment

One representative sequence was selected per cluster, and these were used as queries to search for copies using COPILIST-NR (see below). These representatives can also be found in the database in the melon sequence project site (https://melonomics.net/).

MITEs

MITEs were identified using a combination of two different softwares: MITE_HUNTER (Han and Wessler 2010) and SUBOTIR (Jordi Payet, unpublished). Both base their predictions on the

evidence of TIRs flanking a short sequence, and which are present in many almost identical copies. They differ in that MITE_HUNTER also generates a consensus sequence based on multiple alignments and uses that consensus to identify all the copies in a family. Many of the predictions overlapped but some were specific to either program, so I took the union of the two sets of predicted elements.

1.2.3 Development of COPILIST-NR (COPy Identifier by LInking Split hiTs) to identify representative's copies

The methods described above allowed us to identify and define as well as possible full copies of the transposons that exist in this genome, but a transposon family will often be formed of some full copies and many fragments or copies with deleted portions. The methods used above do not permit us to identify these fragments, namely, LTR_FINDER bases its prediction on the presence of direct repeats that could be LTRs, and we also select the elements potentially coding for retrotransposon-related proteins, therefore only elements with well conserved LTRs and coding region can be detected. Also, the alignment-based approach of identifying full length DNA transposons is dependent on the fact that at least two full-length elements of that family exist. Therefore, to get a comprehensive view of the transposon landscape in this genome, I needed to identify truncated copies of the full-length representatives, and resolve redundancy in the case that a particular genomic region was picked up by various representatives. The most straightforward approach is to use BLAST to identify sequences similar to the representatives, however, copies within a family can vary according to mutations accumulated over time and so a copy will often be composed of fractionated BLAST high scoring pairs (HSPs) (blastn, e-value < e-10). The idea is to assemble these fractionated hits into a copy that spans the greatest length possible of the query. This is not exactly straightforward for various reasons, due to the repetitive nature of the structure of certain TEs, the fact that there can be rearrangements within an element, and the fact that a certain element can be found in many locations in the genome. For this I developed a set of programs, COPILIST and COPILIST-NR, to identify copies of a query sequence by assembling BLAST HSPs.

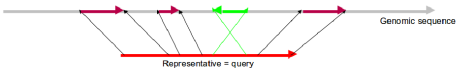
Linking HSPs into copies

Certain criteria must be applied in assembling HSPs together to form a copy, namely that the HSPs must be ordered along the query, be on the same strand, and be separated by at most a certain gap threshold. (**Fig 1.4** – COPLILIST methods) In addition to these criteria, one wants to assemble HSPs such that one finds the set of longest non-overlapping copies of a given element. To solve this optimization problem I represented it as a directed acyclic graph, with as nodes the HSPs which are connected by a directed edge if the previously mentioned criteria are fulfilled. Thus finding the set of non-overlapping copies of an element reduces to finding the set of non-overlapping longest paths in this graph (http://en.wikipedia.org/wiki/Longest_path_problem).

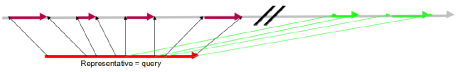
COPILIST - NR
COpy Identifier by LInking Split HiTs
(Non - Redundant)

1) assemble fragmented BLAST hits into a continuous copy

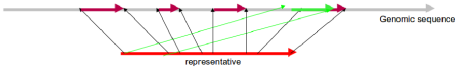
criteria:



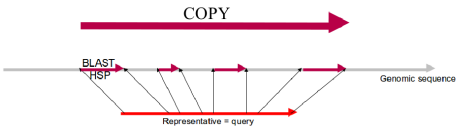
- on the same strand



- within a certain gap range



- ordered along the query



→ represent the problem as a directed acyclic graph, with HSP endpoints as nodes connected by an edge weighted by distance on the query, then connected by an edge if they fulfill the criteria

→ assembling optimal copies is reduced to finding the set of longest paths in this graph

2) resolve redundancy

- pick the “better” of the overlapping copies, either the longest or the one that accounts for the higher fraction of its respective query

- truncate the other one to its closest HSP

Figure 1.4 COPILIST - NR illustration

Resolving redundancy

Though I selected representative sequences to be at least 20% different from each other, within a particular superfamily they remain similar to a certain extent. Thus a genomic region can be identified as a copy of more than one representative. To resolve this redundancy I chose to maintain the longest copy, and truncate any overlapping copies. This is done recursively till there are no more overlaps. (**Figure 1.4** – resolve redundancy)

Information included in the copy annotations

The annotation of each copy carries the following information as tags in the 9th column of the gff3 annotation:

- start and end coordinate with respect to the query (this was particularity important for the subsequent analyses of LTR retro fragments)
- percent similarity to the query, calculated as the average of the percent similarity over it's HSPs
- percent coverage of the query, calculated as the aligned length over total length
- number of complete (> 70% coverage) and fragmented (<70% coverage) elements in the family
- unique ID
- family and superfamily

The SO term for the third column was chosen as “DNA_transposon” for DNA transposons and MITEs, and “retrotransposon” for LTR and non-LTR retrotransposons

I decided to implement my own copy-finding program because the one found in the MITE-Hunter suite

(Han and Wessler 2003) does not supply the genomic coordinates of the copies it returns, nor does it allow parameters to set maximum gap length allowed, nor does it resolve redundancy. For this reason it was sufficient for the purpose of identifying copies of the LTR retrotransposon sequences for refining by alignment, but I found the need to write our own tool as the analysis progressed.

1.2.4 Identification of transposon-related fragments

In order to retrieve degenerate fragments of elements that were not retrieved with any representative, we performed a BLAST search using transposase and retrotransposase protein queries, collected from the NCBI database, excluding any “putative” or “hypothetical” annotations. Regions showing similarity (e < 1e-10) were annotated as transposon_fragment and retrotransposon_fragment, respectively, but not further categorized.

1.3 Subsequent Biological Analyses

Description of the transposon landscape in melon

In melon I identified by homology and structure-based methods 323 transposable element representatives belonging to the major superfamilies of elements previously described in plants. These representatives were used as queries to annotate 73,787 copies in the melon assembled genome, totalling 19.7 % of the genome space. This percentage is similar to that reported for genomes of similar size such as the recently described genome of cacao (Argout et al. 2011). However, this percentage is probably an underestimation of the genome fraction that transposons occupy in melon due to the high stringency of our searches. We opted for a quality annotation and restricted the sequences used in homology-based searches to those that are well-characterized, in order to minimize propagation of errors and uncertainties in the public databases. Thus this data can be used beyond just masking, and to carefully describe the transposon landscape of melon. The overall percentages of sequence covered by superfamily of transposons is summarized in **Table 1.1**.

type	superfamily	% of genome
LTR retrotransposons	copia	5.5
	gypsy	7.2
non-LTR retrotransposons		0.1
retrotransposon related sequences		1.9
Total retrotransposons		**14.7**
DNA transposons	CACTA	1.6
	hAT	0.1
	Mariner	0.1
	MULE	1.9
	PIF	0.3
	Helitron	0.06
transposon related sequences		0.8
Total DNA transposons		**5**

Table 1.1: Transposon content of the melon genome

The retrotransposon elements described account for 14.7 % of the genome whereas DNA transposons represent 5.0% (**Table 1.1**). A total of 87% of the annotated transposon-related sequences were attributed to a particular superfamily of elements and further classified into families. The fact that such a large part of the TEs annotated are classified into families attests to the accuracy of our methods, since very little sequence was left as uncategorized transposon fragments.

Dating LTR retrotransposon insertions

The two LTRs of a retrotransposon are identical upon insertion and accumulate mutations over time. Therefore the degree of similarity between them is an indication of how recently they were inserted, and LTR insertion times can be estimated by comparing the two LTRs (see for example Choulet et al. 2010). I used this strategy to date the insertion time of all LTR-retrotransposons with two at least partially intact LTRs by intra-element comparison of LTRs (see Materials and Methods chapter for more details on parameters and molecular clock used). This analysis showed that, while different families had distinct patterns of amplification over time (**Figure 1.5 B**), most retrotransposons have been inserted during recent melon evolution with a peak of activity around 2 million years (Myrs) ago (**Figure 1.5 A**).

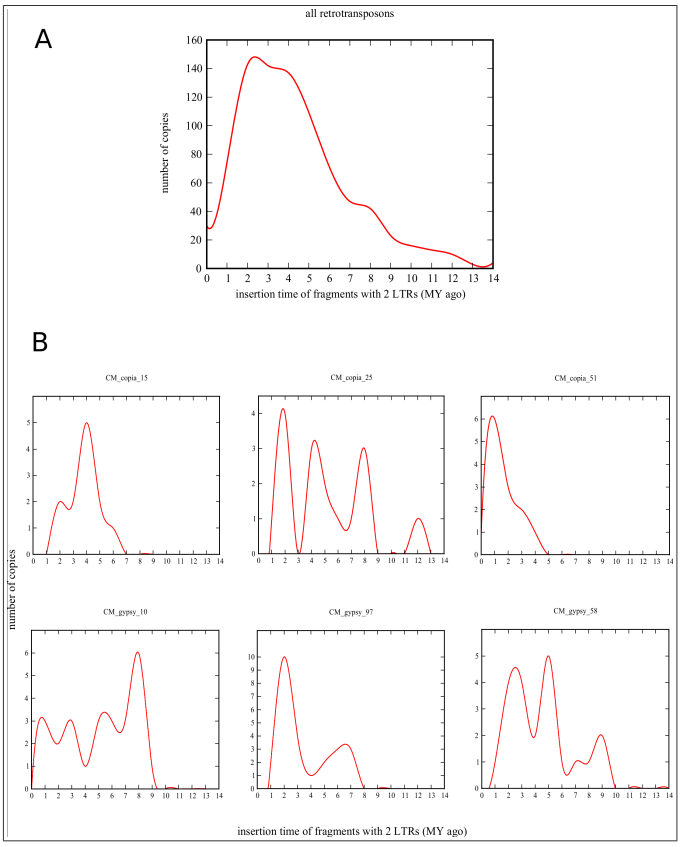


Figure 1.5: LTR insertion dates as calculated by intra-element LTR comparisons

A) all the retrotransposons with two intact LTRs and B) selected examples of individual families

Comparison with the cucumber TE landscape

As cucumber is a close relative, and it's genome had recently been completed, we decided to compare the melon transposon landscape with that of cucumber in order to gain insight into how the melon TEs have evolved since the divergence of these two lineages. Consistent with the fact that we found melon TEs to be recent, no elements homologous to those identified in melon were found in the cucumber genome (data not shown), whose ancestor is supposed to have diverged from melon 10.1 Myrs ago (Sebastian et al 2010). The retrotransposon content of the cucumber genome has been reported as 12.16% (10.43% of LTR retrotransposons) (Huang et al. 2009), which is lower than the value in melon, suggesting that LTR retrotransposon activity has been higher in the melon lineage. As the different approaches used to annotate LTR retrotransposons in cucumber and melon may influence the data obtained, I decided to perform a similarity-based search in the two genomes to estimate relative retrotransposon quantities, providing directly comparable data. I used as queries the most conserved coding regions of retrotransposons, and retrieved twice as much sequence in melon as in cucumber (2.01 v.s. 0.99 percent, respectively), confirming the greater accumulation of retrotransposons in the melon lineage.

The number of sequences, and corresponding genome fraction, related to DNA transposons reported here are substantially higher than in cucumber. Indeed, while DNA transposons have been described to account for 1.24% of the cucumber genome (Huang et al 2009), I report here that 5.0 % of the melon genome is composed of DNA transposons. In order to rule out a methodological reason for this difference, I applied the pipeline used to identify DNA transposons in melon to the cucumber genome, looking for the three most represented superfamilies in melon and cucumber (i.e. CACTA, MULE and PIF/Harbinger) (Table 2, Huang et al 2009, **Table 1.1**). Our results show that all these three families have been amplified in the melon lineage, where the genome fractions each occupy are several fold larger than in cucumber (10X for CACTA, 47X for MULE, and 3.8X for PIF). (**Table 1.2**)

		CUCUMBER		MELON			
superfamily	% query length	# copies	% genome	# copies	% genome	% genome fold difference (melon/cucumber)	# copies fold difference (melon/cucumber)
CACTA	90	13	0.015	116	0.102		
	20-90	169	0.102	1756	0.991		
	0-20	374	0.046	5114	0.536	x 10	x 12,5
MULE	90	3	0.002	112	0.102		
	20-90	73	0.032	1947	1.132		
	0-20	66	0.007	7681	0.702	x 47	x 68
PIF	90	23	0.027	10	0.016		
	20-90	55	0.034	456	0.18		
	0-20	215	0.02	2524	0.116	x 3,8	x 10
unclassified			0.74		0.83	x 1.1	
TOTAL			1.028		4.707	x 4,6	

Table 1.2: comparison of major DNA TE families in melon and cucumber

While comparing coverage data gives us an idea of the transposon content of each genome, it does not reveal anything regarding the evolution of these elements. Indeed, this difference could either be due to an expansion of melon-specific families, or removal from the cucumber genome of elements that had been present in the common ancestor. In order to gain insight into the evolution of the transposon families in these two genomes, we performed a phylogenetic analysis based on the protein-coding sequences of the major DNA transposon superfamilies: CACTA, PIF and MULE. These trees show that for every superfamily investigated, most clades are melon-specific, meaning that the difference in copy numbers is due to an expansion within the melon lineage (**Figure 1.6**). The clades defined mostly concur with the family definitions, though sometimes a given clade corresponded to more than one family (not shown). This confirms the accuracy of our family definitions and that at worst our criteria for defining families were on the stringent side.

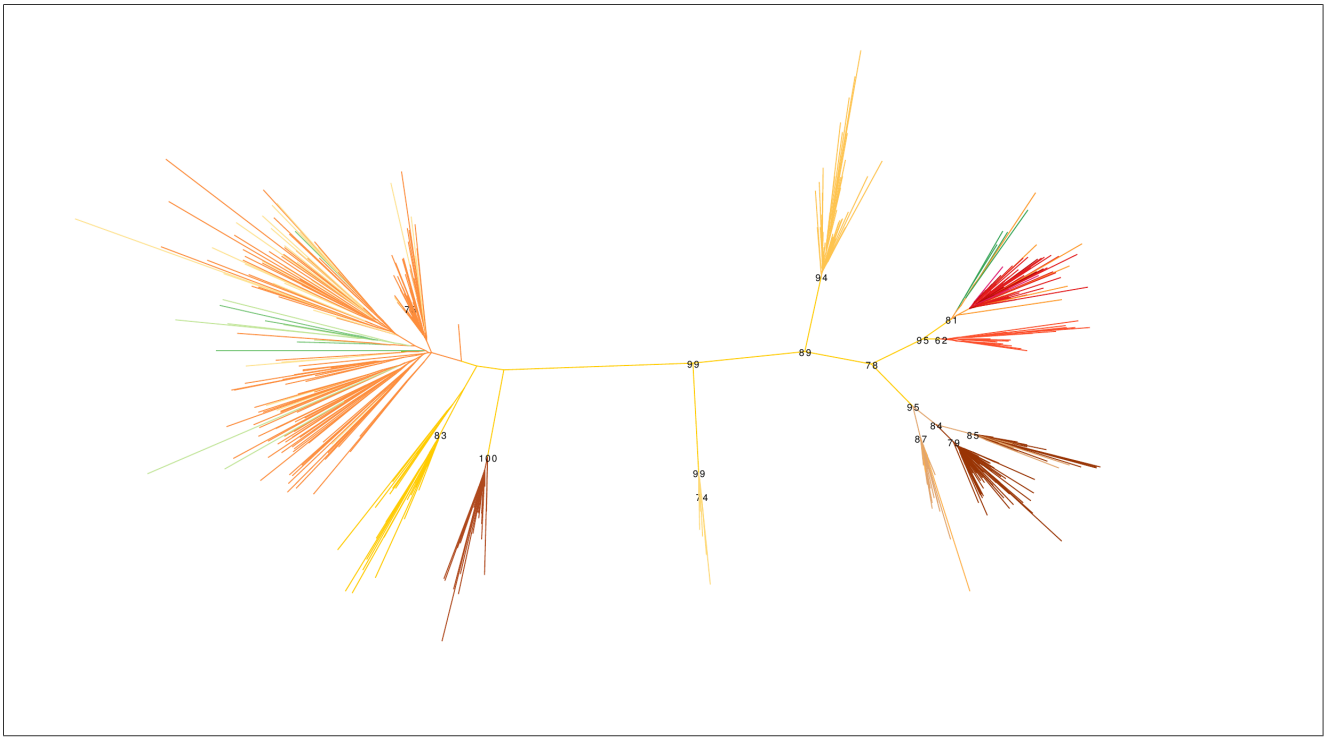


Figure 1.6: phylogeny of CACTA elements in melon and cucumber.

Branches for either species are colored in yellow-ocher and green tones, respectively. Few clades contain a mix of elements from the two lineages, most are melon-specific. Numbers at nodes represent bootstrap values over 100.

As further comparison of the transposon landscapes in melon and cucumber, we compared the spatial distribution of TEs in two collinear chromosomes: chromosome 1 in melon, and chromosome 7 in cucumber. (**Figure 1.7**) While these two chromosomes show syntenic blocks throughout their length (colored plots), the region highlighted in red in melon has been highly expanded. In both chromosomes you can observe an anti-correlation between gene and TE densities, and the expanded region in melon is TE-dense, while the gene-dense region is similar in both. This supports the hypothesis that transposon activity is responsible for the difference in genome size between these two species.

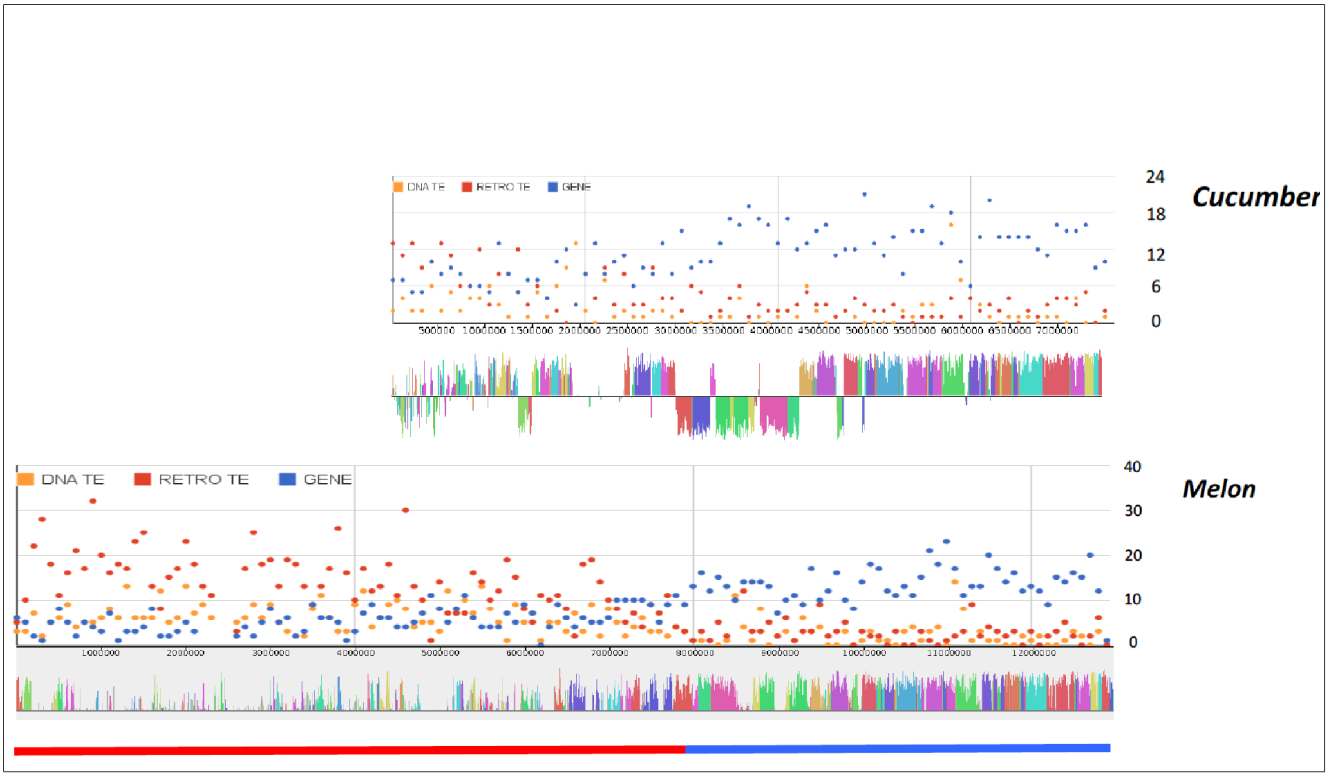


Figure 1.7: distribution of TEs and genes in melon chromosome 1 and cucumber chromosome 7.

Region defined in red has been expanded and shows preferential insertion of TEs

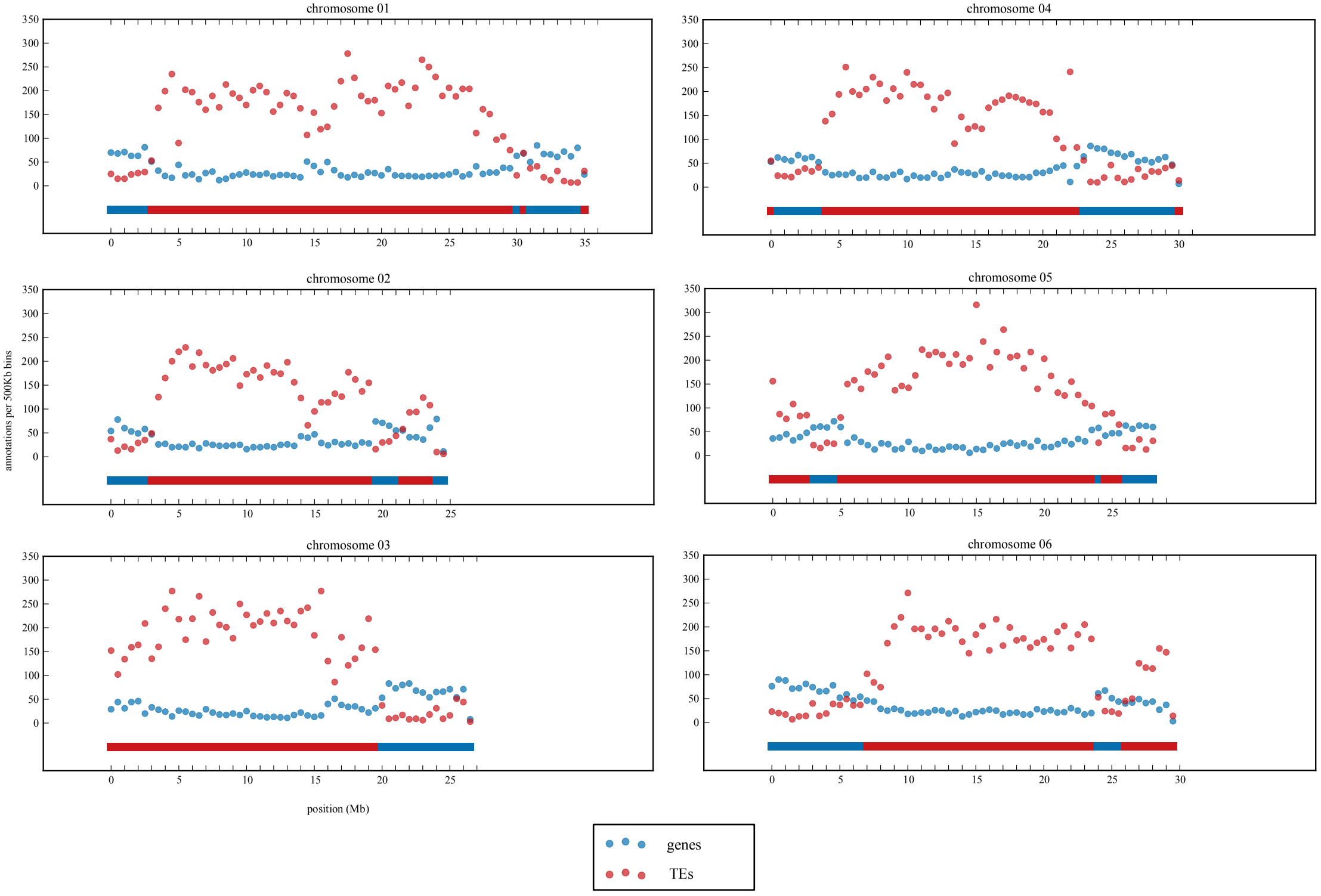
Chromosomal distribution of TEs

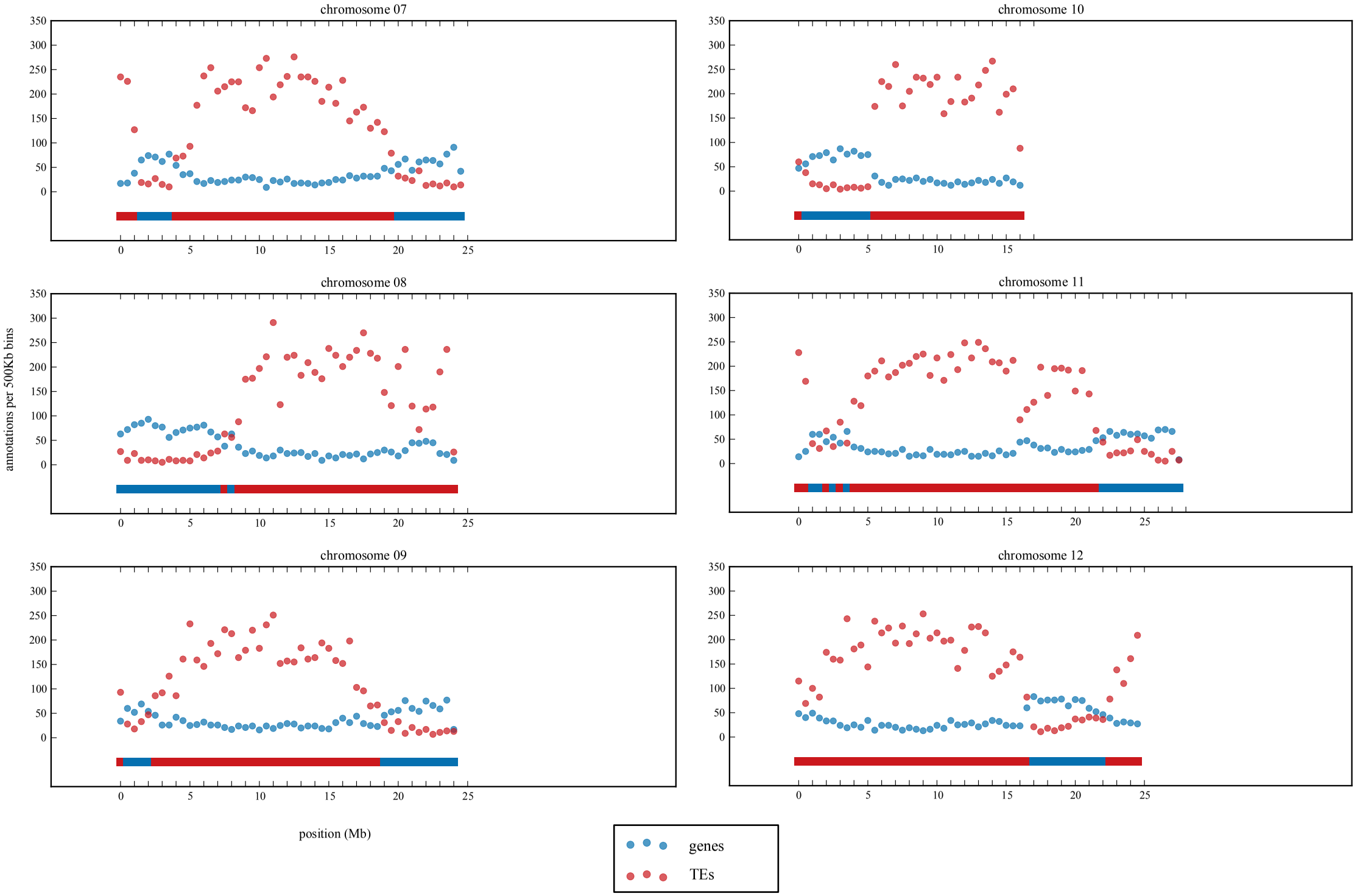
In order to visualize the distribution of TEs along chromosomes and compare it to that of genes, I plotted their respective densities (number of annotations per 500kb bin) for each chromosome (**Figure 1.8**).

next two pages:

Figure 1.8: Chromosomal distribution of TEs and genes for each of the 12 chromosomes in the melon genome.

Colored bars represent gene- and transposon-rich regions represented in blue and red, respectively. A region is considered transposon-rich when the ratio of TEs to genes is > 1, gene-rich otherwise.


chromosome 01
chromosome 04
chromosome 02
chromosome 05
chromosome 03
chromosome 06
annotations per 500Kb bins
position (Mb)
genes
TEs


chromosome 07
chromosome 10
chromosome 08
chromosome 11
chromosome 09
chromosome 12
annotations per 500Kb bins
position (Mb)
genes
TEs

One can observe a striking anti-correlation between the gene and transposon distributions in all the chromosomes, as what was observed in the portion of chromosome 1 analyzed earlier: TE dense regions are significantly depleted in genes. Thus I defined two types of sequence region based on the ratio of TE density over gene density. Ratios greater than 1 are defined as "transposon-rich" regions and ratios less than 1 are defined as "gene-rich". The chromosomal distribution of these two types of regions follows that of recombination rate (Garcia-Mas et al. 2012) and concurs with cytogenetic data (Jordi Garcia-Mas and collaborators, unpublished) indicating that the TE rich regions correspond to pericentromeric regions. This is to be expected as TEs tend to be eliminated at a lesser rate in heterochromatic regions than euchromatic regions (Hollister and Gaut 2009).

Dynamics of LTR retro expansion and contraction

The annotation of a genome is a static picture of the state of the mobilome in this particular individual, but there are characteristics of the elements that reveal their history: for example, phylogenetic analysis showed that the DNA transposons in melon are specific to its lineage, and the intra-LTR comparisons enabled us to date their insertion times. These analyses tell us about the expansion patterns of these families, but the state of the transposon landscape is the result of both expansion and contraction (Devos, Brown, and Bennetzen 2002). Indeed, given the potentially exponential replication of retrotransposons, genomes would quickly become bloated if they were not also eliminated. Retrotransposons are eliminated by small deletions, and also by recombination between their two LTRs (C Vitte and Panaud 2005). Since the LTRs are largely similar, these can be used as substrates for illegitimate recombination, effectively eliminating the internal sequence and leaving a "solo LTR". Analyzing the types of copies within a family – elements with two intact LTRs and some internal sequence, solo LTRs or deleted copies with only one LTR and some internal sequence (**Figure 1.9**)– tell us about the rate and mechanisms of removal that have weighed upon it. In order to evaluate what are the removal forces at work on LTR retrotransposons in the melon genome evolution, I tallied the types of fragments overall for *gypsy* and *copia* superfamiles, as well as for individual families. (**Table 1.2**)

Overall, LTR retrotransposons seem to be eliminated more frequently by deletion than by illegitimate

recombination, though this trend is less marked for *copia* families. In any case, LTR retrotransposon population is kept strictly in check, since only 11% are potentially complete elements.

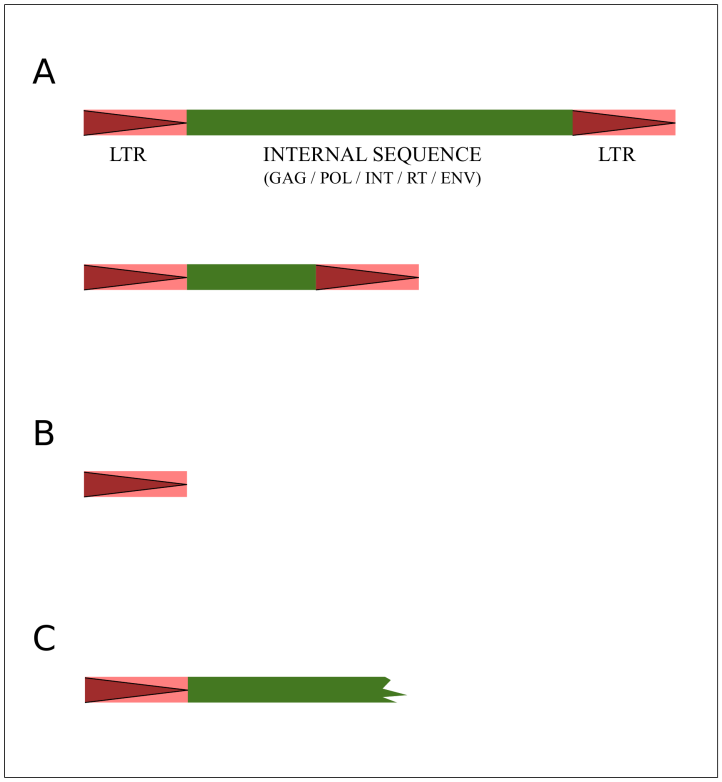


Figure 1.9: types of LTR retrotransposon copies

A) 2LTR_IN: elements with two LTRs and some portion of internal sequence. B) solo_LTR: element who's two LTRs have recombined, eliminating the internal sequence. C) LTR_IN: deleted element in which remains an intact LTR and some internal sequence

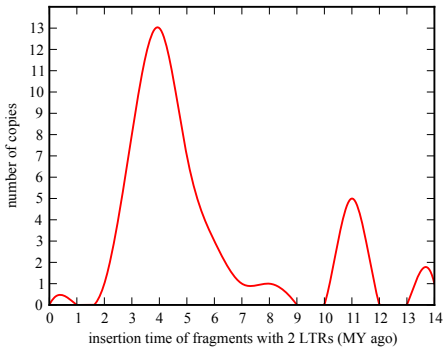
	% copies		
	LTR-IN-LTR	solo LTR	LTR-IN
gypsy	9.86	36.43	53.70
copia	13.51	40.31	46.19
total	11.44	38.11	50.44

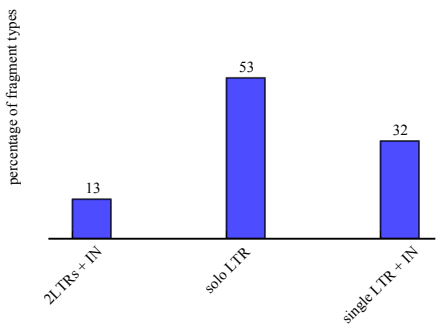
Table 1.2: percent of each copy type for gypsies, copias and overall.

Copias tend to have more elements with 2 LTRs as well as solo-LTRs, while gypsies have more deletion derivatives

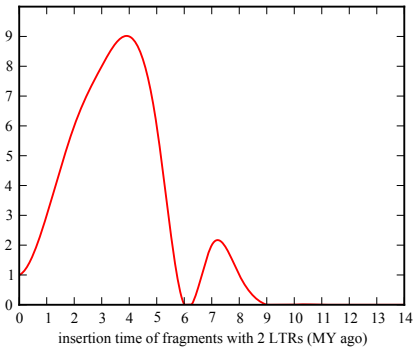
next page: Figure 1.9: insertion dating and proportion of fragment types trace the history of individual families of LTR retrotransposons

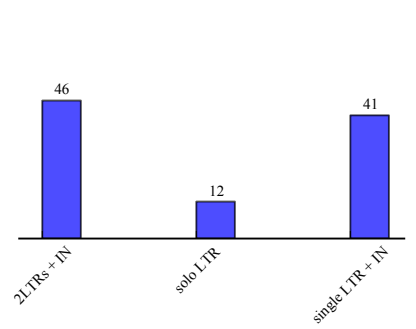
CM_copia_39



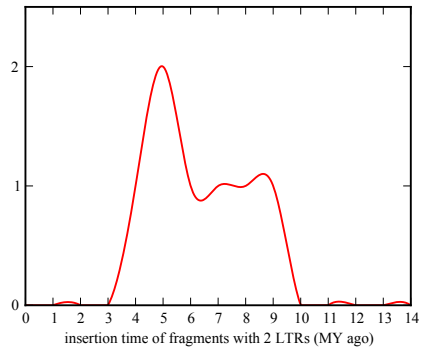


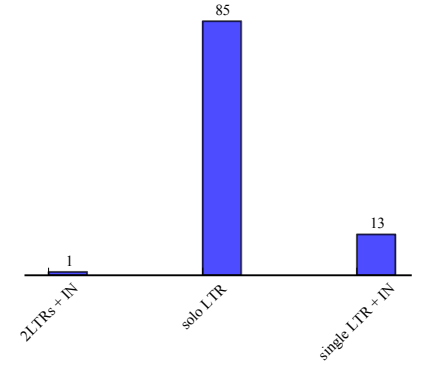
CM_copia_47



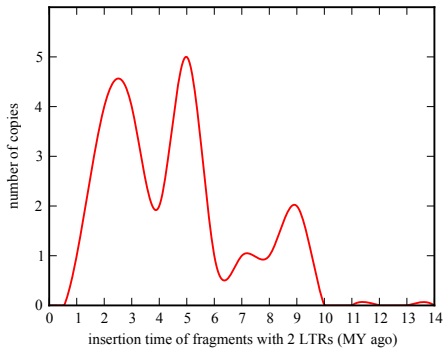


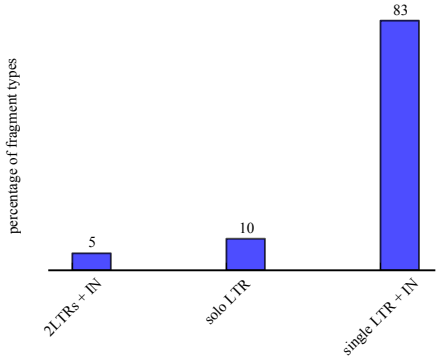
CM_copia_33



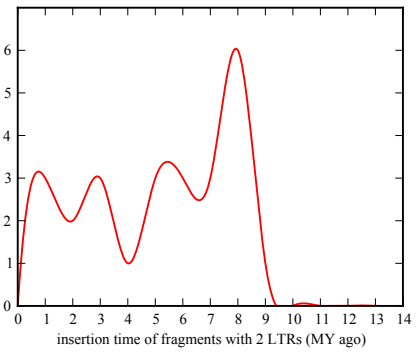


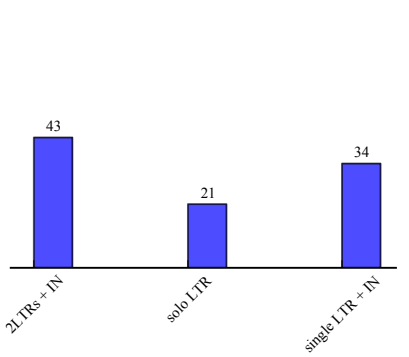
CM_gypsy_58



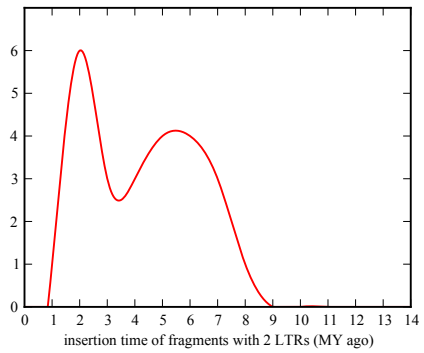


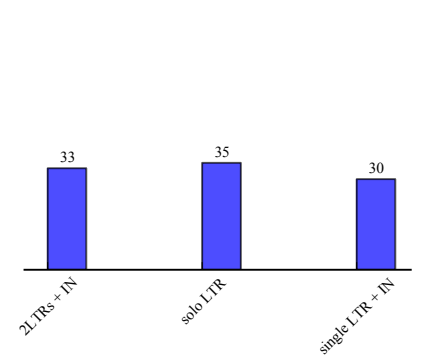
CM_gypsy_10





CM_gypsy_57





1.4 Discussion

The results presented here show that both DNA transposons and retrotransposons have accumulated to a greater extent in melon compared to cucumber, and suggest that transposable elements have played a major role in shaping the melon genome in recent evolution.

The transposons in melon are recent and specific to its lineage

These data taken together show us that the transposons present at this time in the melon genome are specific to its lineage, and have witnessed an expansion in recent evolution. Indeed, there are not only more TEs in melon than in its close relative cucumber, but phylogenetic analysis shows us that most families are not common to both. Had the difference in coverage been due to a loss of copies in the cucumber genome, we would have seen similarities in the families present today, just a difference in number of elements. A recent expansion is consistent with the dating of the LTR retrotransposon insertions, which all postdate the melon-cucumber split, and is likely partially responsible for the difference in genome size between these two species.

LTR retrotransposons are been actively removed from the genome

The analysis of LTR fragments shows us that retrotransposons have suffered constant DNA removal. Interestingly it seems like LTR retrotransposon activity peaked around 2 MYA and that there has been less activity in the very recent evolution of this genome, unlike *Medicago truncatula* (Wang and Liu 2008) where most insertions are within 1MYA and none older than 3MYA are detectable. Alternatively, there recently has been a much more vigorous elimination of LTR retros. The fact that LTR retrotransposons seem to be eliminated more frequently by deletion than by illegitimate recombination is different than in *Arabidopsis* where the two mechanisms operate equally (Devos, Brown, and Bennetzen 2002). This might indicate a better capacity of the latter for elimination, consistent with its compact size. Interestingly in *Arabidopsis* 46% of the LTR retros have 2 intact LTRs, considerably more than the 11.4% melon and which correlates to their overall more recent activity

(Devos, Brown, and Bennetzen 2002).

The bias in fragment types observed for the *copia* and *gypsy* superfamilies is consistent with the fact that gypsies tend to insert preferentially into heterochromatic regions, which have lower recombination rates, and even into other *gypsies*, thus fragmenting them. *Copias* on the other hand tend to insert into euchromatic regions where recombination rates are higher. These are general trends but the history of each family can be quite different, implying that there are features of the individual families than influence removal rates. For example, elements with longer LTRs (e.g CM_copia_33, **Figure 1.9**) tend to form more soloLTRs, consistent with what has been observed in rice (Clémentine Vitte, Panaud, and Quesneville 2007).

TEs accumulate in centromeres

The chromosomal distribution of TEs and their anti-correlation with genes attests to the equilibrium between the colonizing force of TEs and genomes' capacities for damage control. Indeed, mobile element insertions are generally deleterious and those inserted close to genes are selected against (Lockton, Ross-Ibarra, and Gaut 2008; Hollister and Gaut 2009). This distribution has been observed in many species and suggests that centromeres are both a “haven” for TEs and that TEs might also be fulfilling a function there.

In order to get a dynamic view of the impact of TEs in recent genome evolution, it would be extremely interesting to be able to see the very recent movements: those that are still not fixed. Especially relevant to their impact on evolution would be to investigate these polymorphisms in varieties under different selective or environmental pressures. This is the question I will address in the following chapter.

Results

Chapter 2

CHAPTER 2: TE MOVEMENT

2.1: Introduction

While the analysis of the TE annotation in the assembled melon reference sequence has provided a rich amount of information on their nature, prevalence and distribution, as well as insight into their history, this analysis remains a static snapshot of a single individual of a species. A comparison with the TEs in cucumber has provided us with a general view of their evolution but it is only possible to compare coverage and type of transposons: the two genomes are distantly related and the quasi-absence of common transposon families means that the traces of those present in their common ancestor have mostly been erased. This is not unexpected even though these two species retain gene synteny to a high degree, as noncoding sequences evolve and are removed much more rapidly than coding sequences (Freeling et al. 2012). In order to understand the dynamics of transposon activity at a shorter timescale, it is necessary to be able to compare closely related genomes, where it becomes meaningful to identify polymorphisms due to the presence / absence of a TE at given loci. A genome-wide map of these polymorphisms, taken with the evolutionary relationships of the genomes in question and additional genomic maps such as gene annotations, gene expression and epigenetic marks, offers the exciting possibility to start deciphering the impact of transposition on gene and genome evolution. The timescale in which to observe this impact depends only on the samples used: the smallest scale one could imagine would be samples from two tissue types of the same organism. In his review, Lisch (2013) stated that one of the great challenges in TE study is getting a grasp – beyond anecdotal examples – of what is TE's impact on evolution, namely how often does transposition have an effect that is selected for. Varieties of domesticated plants offer a unique system in which to study this, as their evolution is recent and has been subjected to the selection for well-defined traits. Polymorphisms have been identified through molecular biology approaches such as Southern Blot, SSAP and PCR in specific individuals or varieties, but the availability of whole-genome sequences offers an unprecedented possibility for studying polymorphisms at a genome-wide scale. When using sequence-based approaches the more the better: the larger the sample size and the more sequence used, the richer the information derived. Several approaches have already been used to identify TE

polymorphisms genome-wide, some of which have been performed in plants.

An assembly-based approach comparing sequences of three rice cultivars (BACs from *Oryza sativa indica* Guangluai 4 and contigs from the strain BGI 93-11, with their syntenic regions in the *japonica* Nipponbare pseudochromosomes) led to the identification of transposon insertion polymorphisms (TIPs) between these varieties (Huang et al. 2008). This revealed that the most abundant polymorphisms were due to Ty3/gypsy and that some DNA transposon families had differential activity in the three varieties considered. The distribution of these TIPs led to the identification of a chromosomal region corresponding to the introgression between Nipponbare and 93-11. These data integrated with EST data showed that TIPs could affect genes in many ways, including the abnormal termination or alternative splicing of transcripts, change of intron size and modification of expression level. Dating insertions unique to Guanglai4 as preceding the domestication of rice supports multiple independent domestications events of *O. sativa*.

Polymorphisms identified by whole-genome alignments have provided insight into the impact of variation in TE sequences on gene expression and sequence variation. Namely, variable TEs tend to be more strongly targeted by siRNAs, evidence that genomes and transposons are in a constant battle to maintain the equilibrium between regulation and genetic diversity (X. Wang, Weigel, and Smith 2013) Indeed, methylated TEs (as measured by siRNA coverage) can affect neighboring gene expression, decreasing it on average (Hollister and Gaut 2009).

Identification of TE polymorphisms based on whole-genome alignment has the advantage of yielding the sequence of the element that is present (or absent) in either genome, and thus enabling sequence comparisons between elements. However, it is limited by the amount and quality of assembled sequence available. Indeed, repetitive sequences are the most difficult to assemble and some assembled genomes might have sufficient quality to study genes, but will be of limited use for TE analysis. The challenge of genome assembly makes this approach less feasible for large numbers of samples, or for highly repetitive genomes.

In the case that one only has access to the assembled sequence of a reference genome, there are various strategies to identify TE insertions in a sample that do not exist in the reference. Transposon insertions have been studied in human using many different methods, including fosmid – based sequencing (Beck et al. 2010) or hemi-specific PCR specific to LINE elements (Ewing and Kazazian

2010) showing there is a high degree of polymorphism of transposon insertions in human populations. The recent advances in paired-end sequencing technology, and its more accessible cost, have led to the generation of large quantities of data that beg this type of comparative analysis. It has even been postulated that next-generation sequencing offers better possibilities for detection of structural variation due to TE movement than other comparative genomics methods: array-based methods are notoriously blind to the "difficult", repetitive regions of the genome, and the challenges of assembly make these approaches unfeasible for large numbers of samples and preclude the identification of heterozygosity. (Alkan, Coe, and Eichler 2011).

Paired-end sequences combined with 454 data have been used to map polymorphic TE sites in human populations (Stewart et al. 2011) or human cancer lines (E. Lee et al. 2012) but unfortunately in the former the authors did not provide the software that implements their approach, and in the latter the tool is specifically designed for the human genome (reference sequence and transposable elements are hard-coded). In their review on TE polymorphism detection methods, (Ray and Batzer 2011) point out that the aforementioned methods are tuned towards detecting human-specific elements (Alus and LINEs), and their extrapolation to other organisms would lead to false negatives. The paired-end detection approach has also been used in certain plant genomes, such as a study by Sabot et al. (2011) which brings to light the evidence of transposition as a result of standard plant breeding practices. They show that different transposons in the same plant are activated during rice cell culture and lead to novel insertions in the resulting cloned line, indicating that the stress of cell culture may have deeper impacts than expected on the genome of plants cloned in this manner. However, the pipeline they developed for this analysis is not supplied.

We have seen that comparative analysis of closely related genomes can yield rich information regarding to the evolution of TEs and their impact on genomes. Comparing closely related genomes allows the identification of TE movement before it is eliminated or mutated beyond recognition. This is important as most TE insertions are usually neutral, and therefore not conserved (Freeling et al. 2012). Even exapted TE functions can become quickly unrecognizable, as for example domesticated MULE transposases that maintain homology at the protein level but have lost the rest of the TE sequence (Joly-Lopez et al. 2012).

In the pursuit of studying the role of TEs in the evolution of the melon genome, we have been lucky to collaborate with the groups of Dr. Puigdomènech and Dr. García-Más in our center who have

sequenced seven melon varieties using Illumina paired-end reads. The varieties considered come from diverse geographical locations (Europe, Asia, India) and have been submitted to varying selective pressures: some are highly domesticated and others are more wild. These seven lines represent a good sample of the phenotypic diversity found in melon, and cover the main phylogenetic groups of the species (see **Table 2.2** in this chapter and Materials and Methods for a more complete description of the varieties used). Our collaborators are using this data to investigate variation between these lines: identifying SNP, small InDels, and other non-TE structural variations. My goal within this project is to identify the polymorphisms due to TE insertions and deletions between these seven varieties, with the objective of investigating the history of TE activity in the evolution of this species, and to what extent it may have impacted its evolution.

When identifying transposon-related structural variations in a sample with respect to the reference, one looks for two types of variations: either “deletions”, which would correspond to an element present in the reference and not in the sample, or “insertions”, being an element present in the sample and not in the reference. However whether a polymorphism is an “insertion” or a “deletion” depends completely on which genome is taken as the reference, and does not make any implications as to the evolutionary process of TE insertions and deletions that might have led to this observation. Indeed, an “insertion” could be either a TE inserted in the sample, or a TE deleted in the reference. (**Figure 2.1**) . Therefore from here on the terms insertion and deletion when used with respect to a reference sequence do not imply anything as to the mode of transposition.

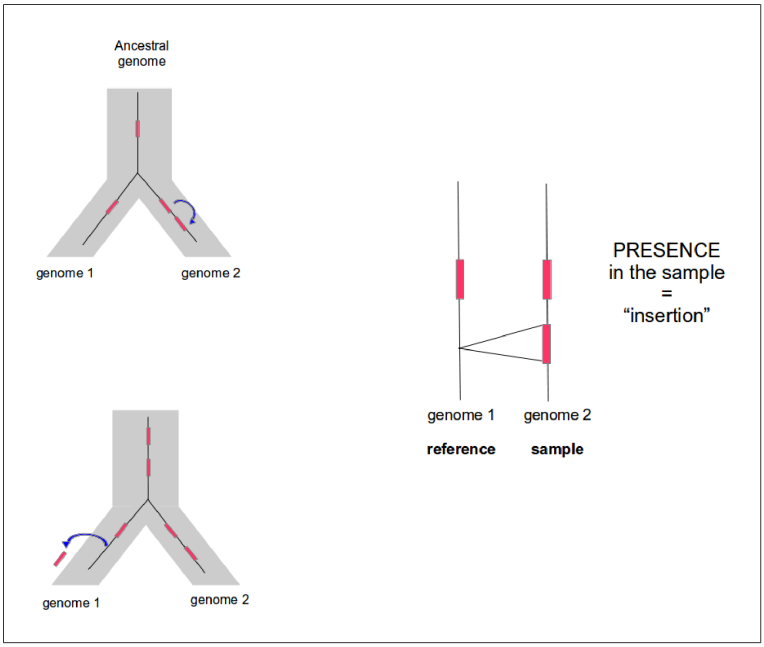


Figure 2.1: different transposition events can lead to the presence of a TE in one genome with respect to another

There already exist several softwares to detect "absences" (such as Pindel (Ye et al. 2009)), and detecting these events is more straightforward in the sense that tools developed for detecting absence of any type of sequence can be directly used for detecting transposon absence in the sample, by taking only the predictions that correspond to a TE annotated in the reference. However detecting "presence" is more challenging, since the differing sequence is divided amongst the sequence reads. There exist some tools for detecting general insertions using paired-end sequencing data such as HITSEQ (Hajirasouliha et al. 2010) which identifies them by assembling unmapped reads with those spanning breakpoints. This approach, however, is not suitable for TEs which are by definition repetitive and difficult to assemble. A similar approach which is oriented to TE insertion discovery is VariationHunter (Hormozdiari et al. 2010) but we found it unsuitable for our pipeline as it is based on an unpublished short read mapper and it requires an artificial chromosome of consensus TE sequences.

For these reasons, I have developed a tool which is designed specifically to detect the presence of a transposon in the sample with respect to th reference, based solely on the paired-end mapping and the annotation of transposable elements in the reference genome.

2.2: Tools developed and used for structural variation detection in whole-genome sequencing

2.2.1 Algorithm principles and design

This software was designed to be fast, easy to use and flexible, as well as provide output that readily lends itself to downstream analyses. It is also designed to be accurate, both in the position of the predicted insertion and its specificity.

It is easy to use because I designed it to only require as input the mapped reads (in bam format, the most commonly used mapping format) and the transposon annotation of the genome. It is flexible in the sense that it can adapt itself to different types of libraries (fragment length and read length) by automatically calculating these from the input bam file. It can also harness the power of multithreaded workstations and perform the bulk of the calculations in parallel, significantly speeding up the runtime. It utilizes softclipped reads which span the actual insertion site and have been mapped only on part of their length, to predict the insertion site down to a 6bp interval as well as predict the allelic ratio of the insertion. Also, I have developed a system for parameter optimization using simulated data which maximizes specificity and sensitivity and that can be extrapolated to any library, thus making these predictions quantifiably reliable.

Finally, the output is both in standard gff3 format (http://www.sequenceontology.org/gff3.shtml) which can be directly loaded into a genome browser as well as in table format which can be easily manipulated with standard unix tools (grep, awk) to extract relevant information. Companion scripts are provided to filter the gff3 file according to various metrics as well as extract and assemble the reads of a given predicted insertion site, yielding the sequences necessary for primer design to verify the predicted insertions (see Materials and Methods)

2.2.2 Algorithm description

This software predicts TE insertions present in the sample based on the fact that read pairs which come from fragments spanning the border of the transposon(**Figure 2.2 B**), when mapped to the reference, will have one read map to a unique genomic location (the “anchor”) , and the other map at a discordant distance to another similar transposon (the “TE mate”) found somewhere else in the reference (**Figure 2.2 C**). This relies on the fact that TEs are found in multiple similar copies throughout the genome, and that the TE present in the sample will likely be similar to another TE in the reference. It is important here to only select reads mapping at a discordant length, as a read pair spanning a TE but mapping properly indicates the presence of a TE at that given location in both the reference and the sample, i.e. no polymorphism. As the DNA fragment these reads come from spans the TE border, the putative insertion has therefore occurred within a fragment length of the mapped position of the anchor read. For a given TE insertion, there will be fragments spanning either border, thus anchors mapping to the forward and reverse strands. Sets of overlapping anchor reads are clustered together on either strand, and a pair of forward and reverse clusters that overlap in their prediction interval are considered to predict a putative TE insertion within that overlap. **(Figure 2.2 D)**.

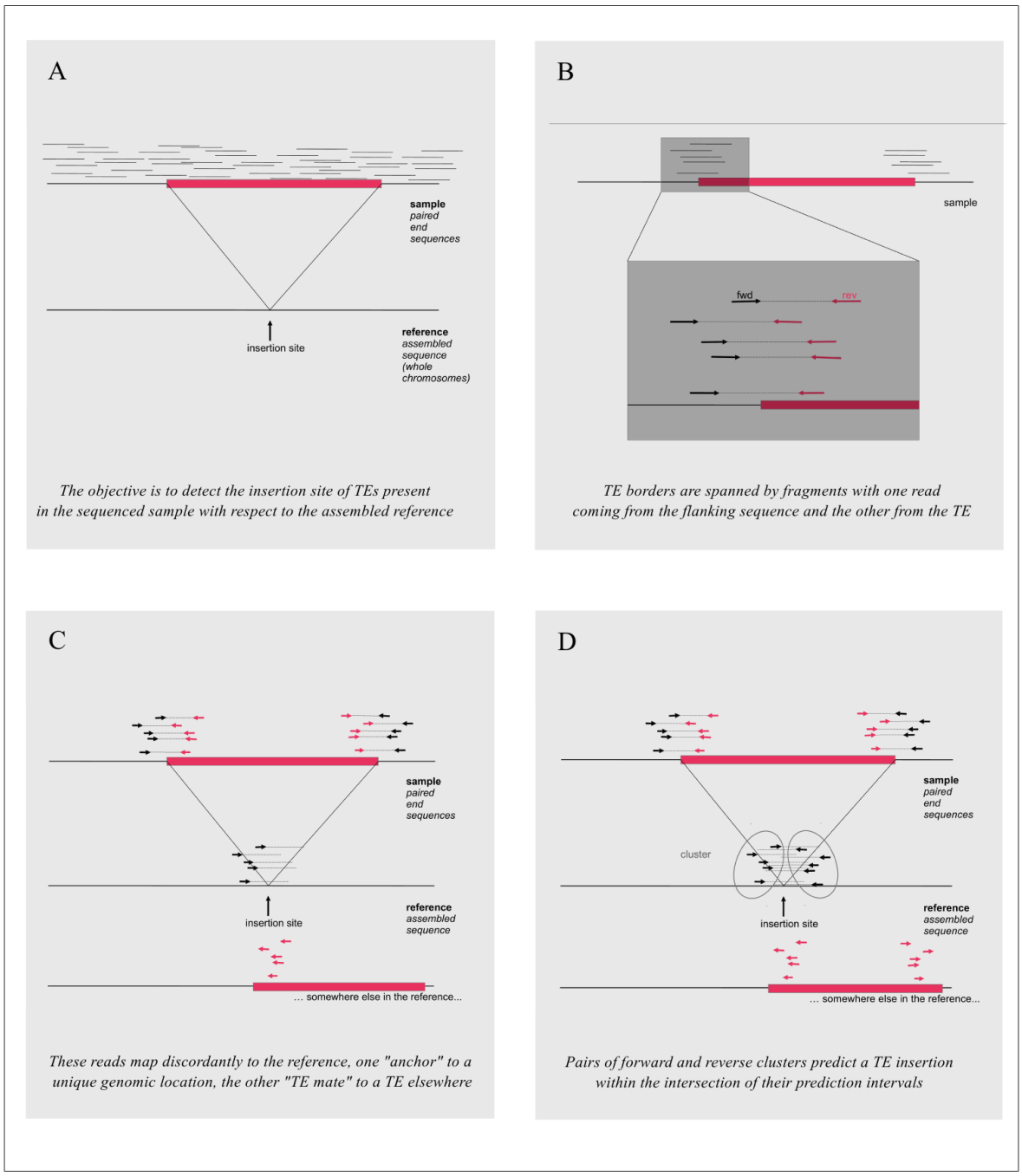


Figure 2.2: Jitterbug algorithm

Since the predictions rely on uniquely mapping anchor reads, insertions in repetitive or otherwise unmappable regions (low-complexity or Ns) are not detectable by this method. Softclipped reads are used to narrow down the prediction interval if possible. A sequenced fragment that spans the insertion site will generate a read pair that maps discordantly, and one of those reads might itself span

the insertion site. When is is mapped, part of the read (which comes from the transposon) will not be mapped and some mapping algorithms (such as bwa) will “clip” it, basically masking the unmappable part of the sequence. The clipped site indicates the exact insertion breakpoint, though depending on the mapping parameters the read might be clipped up to 3 or 4 basepairs after the breakpoint. Several reads that are clipped at the same site support that this exact site is where the breakpoint occurred (**Figure 2.3)**

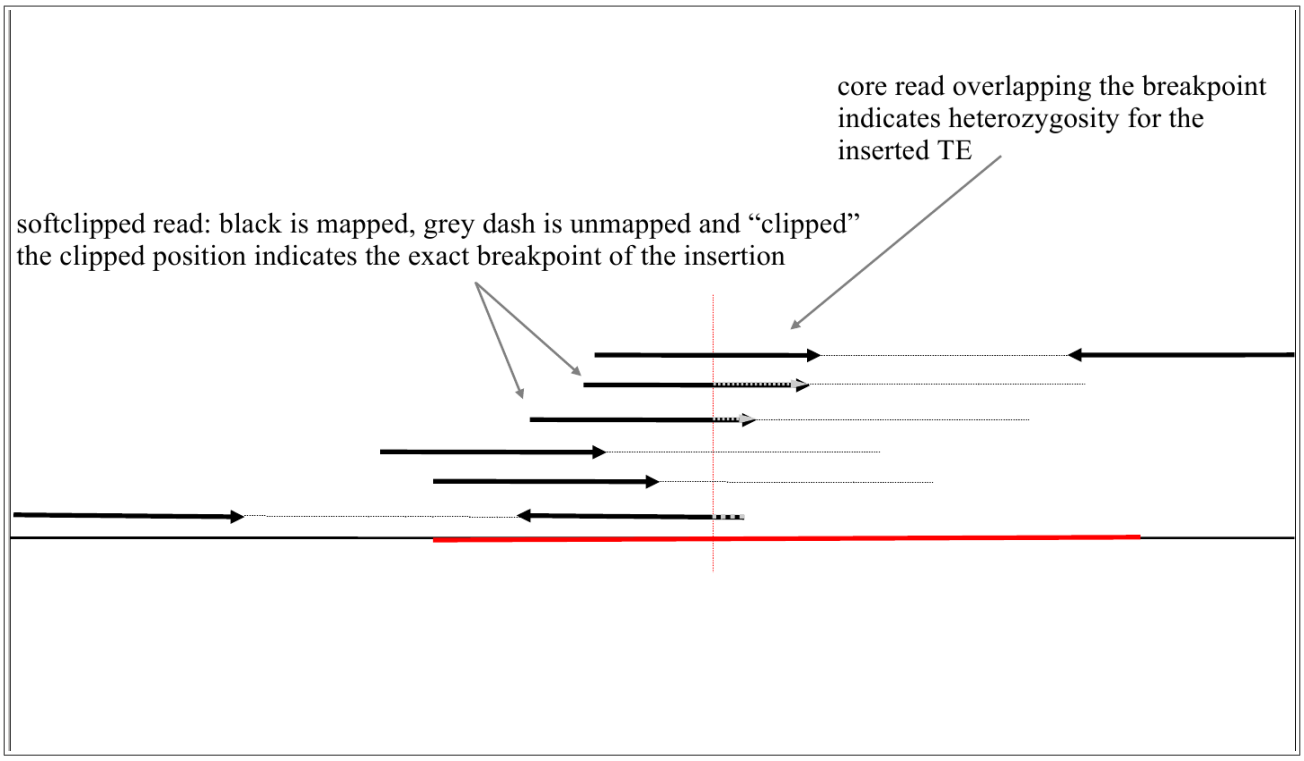


Figure 2.3: Softclipped reads predict the exact insertion position

If the exact insertion site is defined, it is possible to determine whether there are any reads that are properly mapped and span the insertion site. These indicate the presence of an “absence” allele, while the softclipped reads indicate the “presence” allele. Indeed, if there is an allele where no TE insertion is present, there will be reads that map properly to the reference at that location. The relative support for these sets of alleles can be calculated as the ratio of softclipped reads to the total number of softclipped and “core” reads. A ratio around 0.5 means half of the reads are softclipped and support a heterozygous state for the insertion at that locus. A ratio near 1 supports homozygous state for the

insertion, and a ratio near 0 supports homozygous state for lack of insertion, meaning the prediction might be a false positive.
Please see Materials and Methods for a detailed description of algorithm implementation, requirements and usage.

2.2.3 Parameter optimization with simulated data

Generation of simulated dataset

As is the case with any computational prediction – which inevitably make assumptions regarding the phenomenon you are trying to predict, the genome, the data – it is important to evaluate the accuracy of our TE polymorphism predictions. One method is to verify these predictions by PCR, designing primers to detect both the presence of the TE as well as the empty site. This we have performed for a subset of the predictions (experiments performed by Cristina Vives, member of our lab, data not shown), but it is not feasible for a large number of predictions, and we designed an *in silico* experiment to test our tool. For this we were lucky to dispose of the perfect dataset: the paired end sequencing data of the reference line DHL92. This sequencing run came from the same individual as the reference genome, and offers the opportunity to design a simulated data set to test the prediction algorithm. I modified the reference genome by "cutting" out a subset of the annotated TEs and "pasting" them into random positions. The absent sites should then be detected as insertions in the resequenced data. The TEs to be shuffled were selected as a random 30% of the entire TE annotation, only considering the elements that represent a significant fraction of their respective representative sequence. The idea behind this is that the TEs that would have moved recently will likely not be overly deleted, and therefore chose elements that cover at least 50% of their representative query's length. This subset covers the range of sizes and superfamilies annotated in the melon genome. This modified genome contains 1871 simulated deletions, that should be detected as insertions by mapping the resequence of the reference. **(Figure 2.4)**

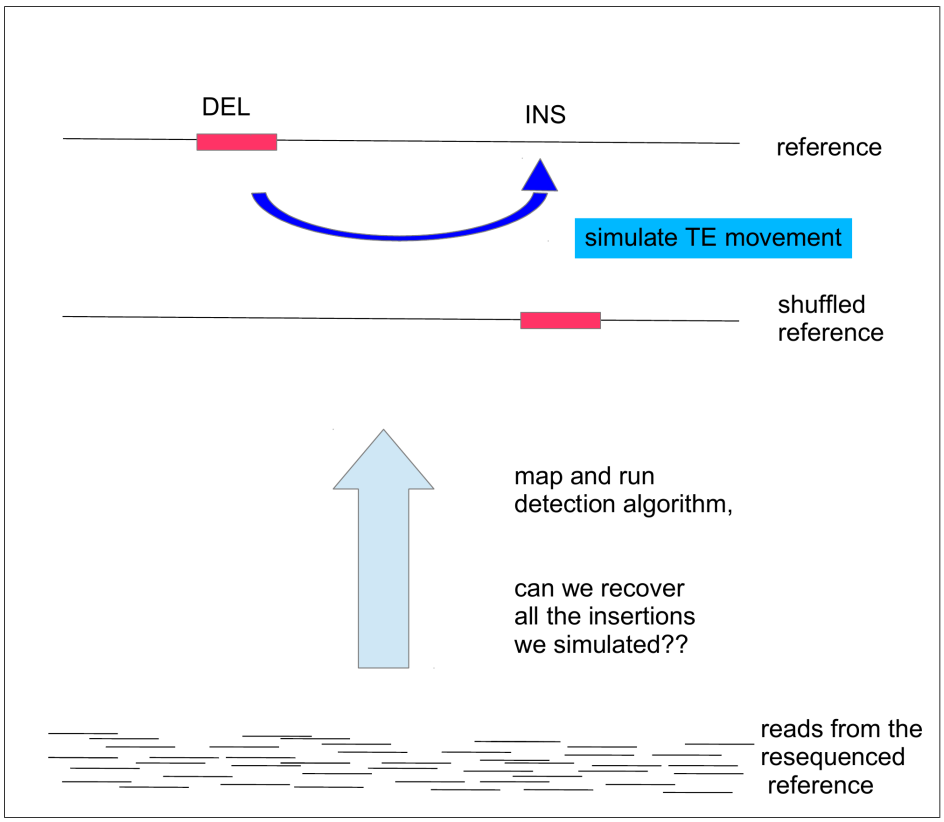


Figure 2.4: simulated data for jitterbug evaluation

As mentioned earlier, insertions in repetitive regions or Ns are not detectable, so any of the simulated deletions within 500bp of an annotated TE or N island were not counted in the calculation of sensitivity.

Evaluation of sensitivity and specificity

Of the 1131 detectable insertions, 1016 were identified, meaning that the detection sensitivity is of 89%. However, of the 3372 total predicted insertions, 1016 were true positives (TP) and 2356 were false (FP). This means that the specificity was poor, at 30%. **(Table 2.1)**

	raw
sensitivity	89.83
specificity	30.07

Table 2.1: sensitivity and specificity of raw jitterbug predictions

Here there are two issues to consider: first, is there some way to differentiate the TP from the FP? and second, what are the possible reasons why some of the insertions are not detected?

In order to address the first issue of specificity, I have established various metrics for the insertion predictions and evaluated their importance in discriminating true positives (TP) from false positives (FP) in the simulations.

The metrics established for evaluating the predictive power of a cluster are illustrated in **Figure 2.4**:

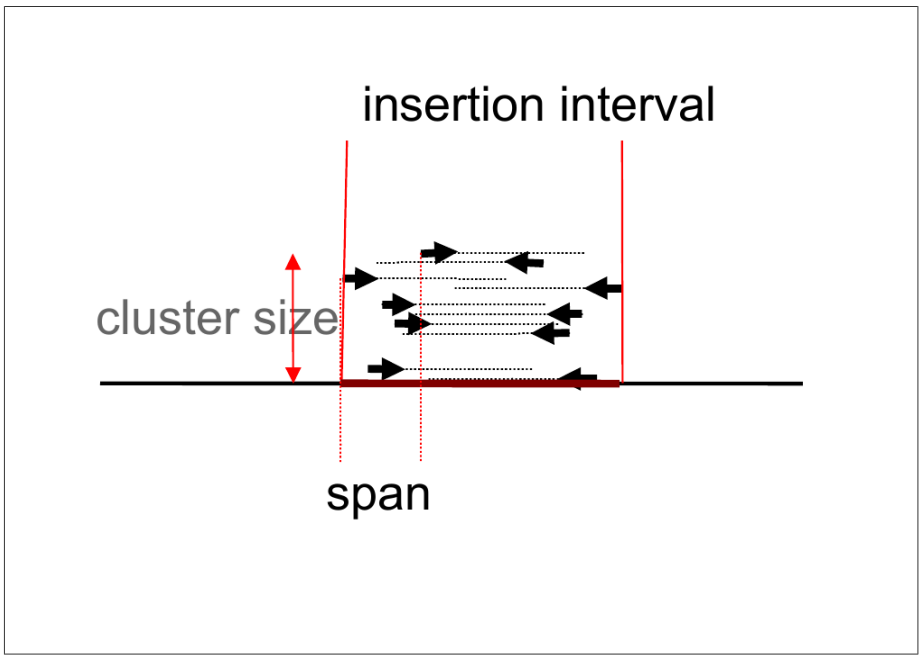


Figure 2.4: metrics for evaluating jitterbug predictions

length of the interval within which the insertion is predicted to have occurred.

cluster size, i.e. the number of reads constituting the forward and reverse clusters which, paired, predict an insertion (in the charts, supporting_fwd_reads and supporting_rev_reads)

span, the distance between that start position of the two most distant reads in a cluster. A span of 0 means the reads are stacked.

Figure 2.5 shows the predictions generated against the simulated genome, plotted according to these metrics. Luckily we see that there are different distributions for TP and FP for each of these metrics, which means that cutoffs for each metric can be applied in order to minimize the FP while maximizing the TP.

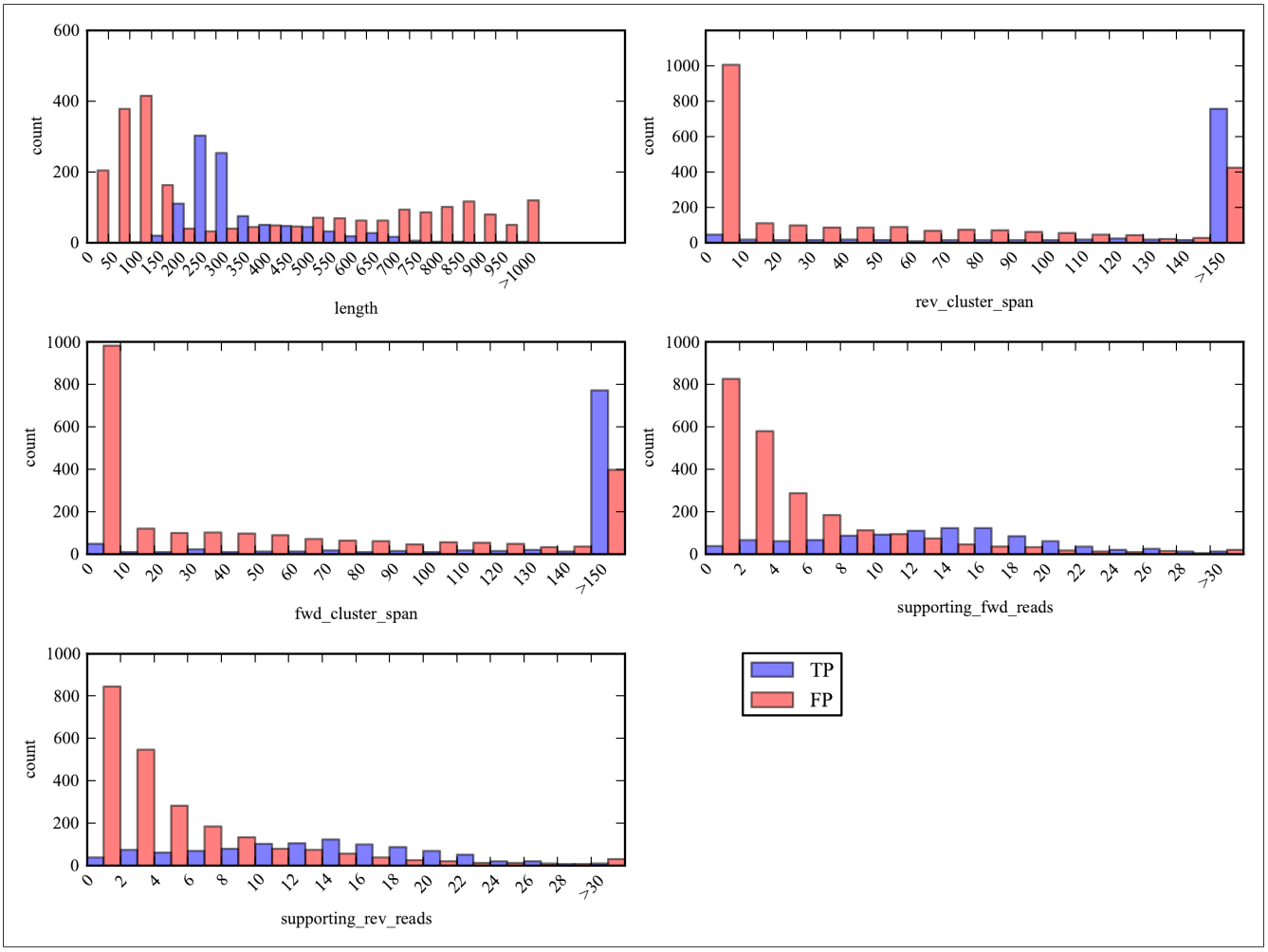


Figure 2.5: plots of raw TP and FP according to different metrics

Filtering criteria to improve performance

After experimenting with various combinations, he cutoffs I settled on are:

150 < length < 700

10 < span

2 < number of reads < 50

With these criteria, the specificity is increased to 82% from 30% while only decreasing the sensitivity to 84% from 89%, and the distributions of TP and FP are no longer distinguishable (**Figure 2.6**). These cutoffs were used for the subsequent analyses in melon.

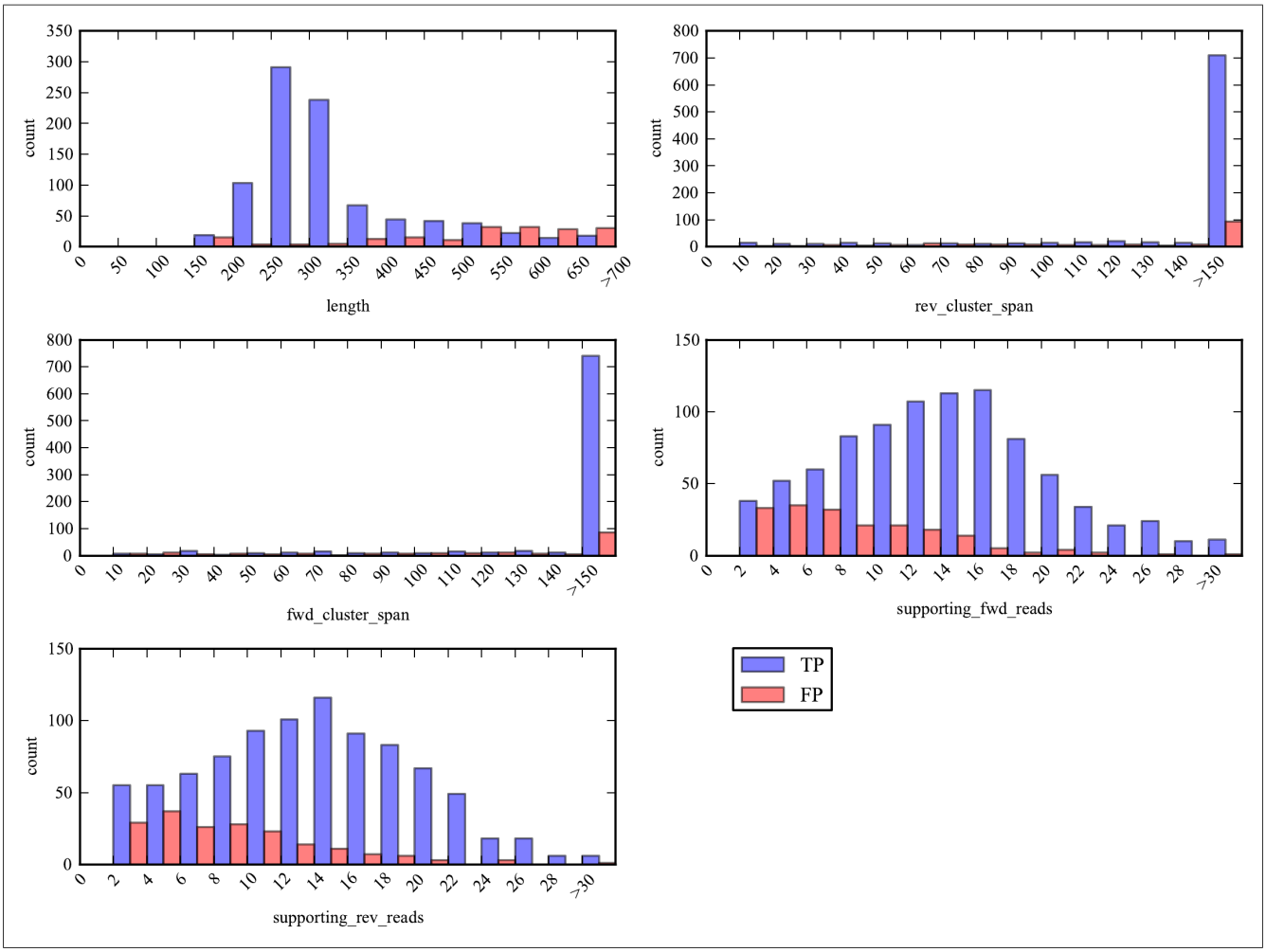


Figure 2.6: plots of filtered TP and FP according to different metrics

That the span be more than 10 eliminates cases of stacked reads, which can arise from PCR duplicates or other sequencing errors, and limiting the coverage to be between 2 and 50 eliminates flukes due to just one read. The length criteria are consistent with the way the algorithm was implemented, and are a function of the characteristics of the sequencing library: 500bp fragments with

150bp reads. It is logical that the insertion interval should be greater than the read length, (smaller would mean the forward and reverse anchors are overlapping) but smaller than two the fragment lengths minus twice the read length (longer would mean the forward and reverse clusters are separated by more than expected within the allowed standard deviation. See Materials and Methods for how fragment length standard deviation is taken into account). Thus these criteria can be extrapolated to any other library, as a function of the fragment and read length.

Evaluation of reasons for undetected predictions

In order to determine whether there was a particular factor that influences the undetection of a prediction we investigated the influence of element length, mapping quality and coverage at the insertion site for the detected and undetected predictions. (**Figure 2.7**). Small elements (less than 100bp) are overrepresented in the undetected elements, even though such small elements can also be detected. We reason that since the detection relies on reads spanning the insertion site, there is a lower probability that there be reads spanning both borders for smaller elements. Mapping quality seems to influence detection as well, as FP are enriched in low-quality mapped reads (less than 30 MQ). Finally, the determinant factor in detection seems to be coverage in that region: indeed, the average coverage of undetected reads is very low. Understandably, an insertion cannot be detected if there are no reads mapping in that location. This can be due to either a lack of sequencing data for that particular region, or unmappability in that region. Low complexity sequences can be a source of either, as they are both hard to sequence and map.

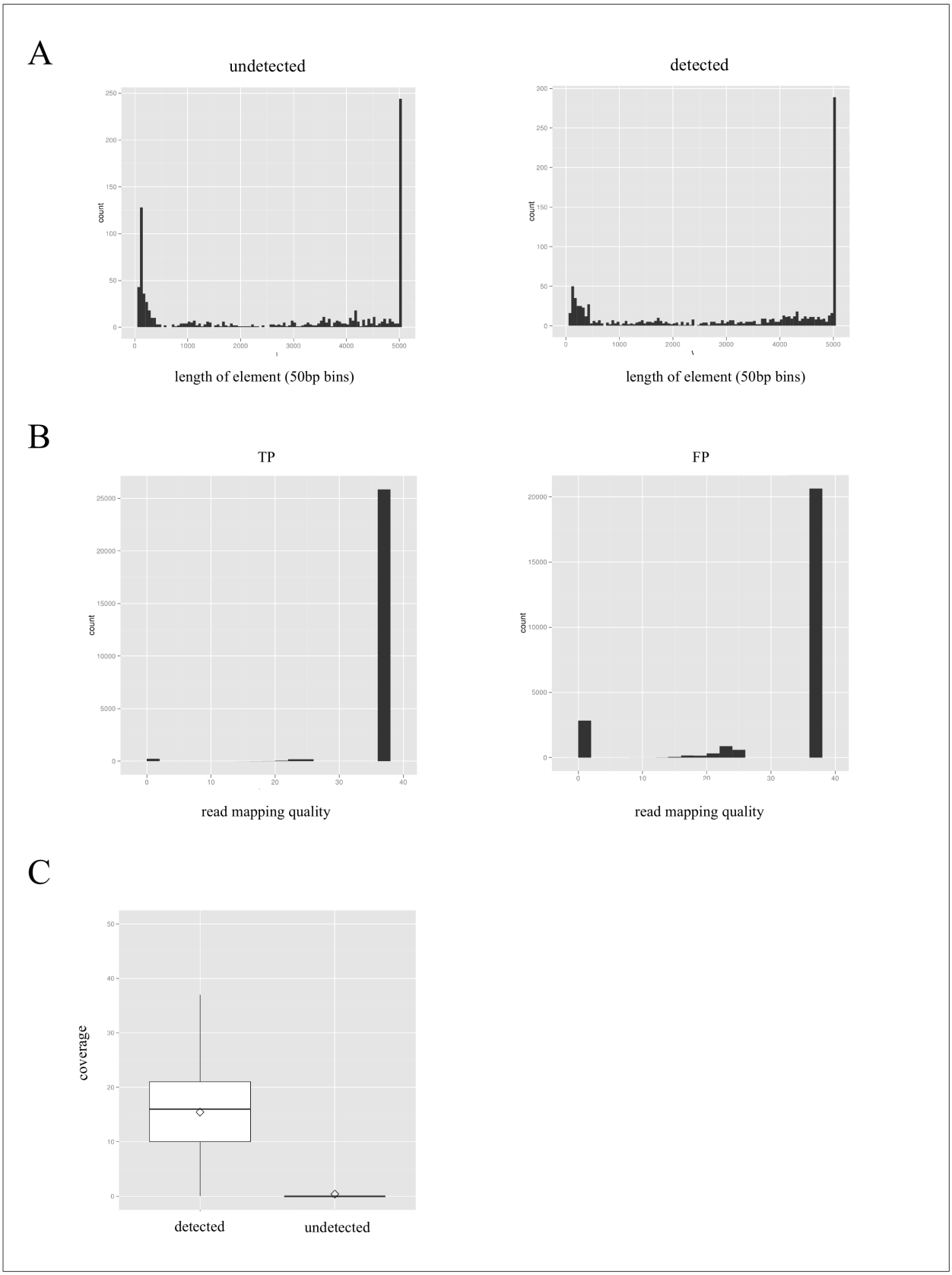


Figure 2.7: factors influencing performance of predictions:

A) length of inserted element B) read mapping quality C) coverage at insertion site

2.3 Biological Analyses

2.3.1 TE polymorphisms in 7 melon lines

Seven melon varieties, from different geographical locations and under different selective pressures, were resequenced by paired-end Illumina sequencing at ~ 17X coverage. (**Table 2.2**). Studying the transposon-related polymorphisms between these lines provides an interesting opportunity to get a more dynamic picture of the role of transposition in the recent evolution of this species' genome.

Plant designation	Accession no	Code	Cultivar group	subspecies	Origin	Reference
DHL92[1]	DHL92	DHL92				Garcia-Mas et al 2012
Piel de sapo	T111	PS	*Inodorus*	*melo*	Spain	Garcia-Mas et al 2012
Songwhan charmi	PI 161375	SC	*Conomon*	*agrestis*	Korea	Garcia-Mas et al 2012
Cabo Verde[2]	C-836	CV		*agrestis*	Cabo Verde	Gonzalez et al 2013
Irak	C-1012	IRK	*Dudaim*	*melo*	Irak	Gonzalez et al 2013
Védrantais		VED	*Cantaloupensis*	*melo*	France	unpublished
Trigonus[2]	Ames 24297	TRI		*agrestis*	India	unpublished
Calcuta	PI 124112	CAL	*Momordica*	*agrestis*	India	unpublished

[1]DHL92 is a doubled haploid line derived from PI 161375 x T111 and represents the melon reference genome.

[2]Unknown cultivar group

Table 2.2: origin of sequenced melon varieties

Using Jitterbug to detect insertions, and Pindel to detect deletions, both with respect to the reference genome, we have established a presence / absence map of polymorphic sites. (See Materials and Methods for more details on the deletions detection with Pindel)

TE polymorphisms detected in the seven melon lines

We have detected insertions and deletions in the seven melon varieties, as well as in DHL92, the resequenced reference (**Table 2.3**). The predictions in DHL92 are very low: 4 deletions and 27 insertions, less than 1% and 5% of the number of predictions in the other lines, respectively. These are

necessarily false positives and the fact that we have so few confirms that our false positive rate is low, any of these that overlapped with predictions in the other seven lines were excluded. The reference sequence is a double haploid line of T111 and PI, thus it is to be expected that the number of predictions in either of these parental lines is lower: roughly half than the other lines. Overall there are more insertion predictions than deletions. There is no biological reason to expect that there would be more differences due to absences in the varieties compared to DHL92, indeed as we mentioned in the introduction these terms are dependent on which genome is taken as reference. What can influence deletion detection is the quality of the assembly: a TE needs to be properly annotated in the reference for its absence to be detected in the sample. Thus, the presence of mis-assembled repetitive regions or N islands that actually contain TEs will impede detection of deleted sequences. Insertions, however, are less likely to be affected by unassembled regions, as by definition they can only be detected in unique genomic regions. We know that the version CM3.5 used has a certain amount of Ns, which can be a reason for the difference in insertion and deletion detection rates.

	deletions	insertions
CV	251	323
IRK	229	428
T111	180	313
PI	165	199
TRI	424	500
CAL	376	489
VED	341	479
DHL92	4	27

Table 2.3: Insertions and deletions detected in the sequenced melon lines

We combined the insertion and deletion predictions in the seven melon lines into a set of polymorphisms: any position where either a deletion or insertion has occurred in any line is

considered polymorphic, and we constructed a map of polymorphic sites (PM sites) annotating the presence or absence of a TE in each line at each polymorphic position, its superfamily and family when possible. (**Table 2.4)**

superfamily	number polymorphic sites	percent polymorphic sites
gypsy	1268	46.43
copia	670	24.53
non_LTR_retro	6	0.22
nonLTRretro	4	0.15
retro_transposon_fragment	83	3.04
total retro	2031	74.37
CACTA	105	3.84
hAT	15	0.55
MULE	349	12.78
Mariner	4	0.15
MITE	15	0.55
helitron	5	0.18
PIF	78	2.86
DNA_transposon_fragment	33	1.21
total DNA transpo	604	22.12
uncategorized	96	3.52
Total	**2731**	100.00

Table 2.4: polymorphic sites in melon varieties

In this manner, we have identified 2731 polymorphic sites across the melon genome, of which 96% have been categorized as corresponding to retroelements or DNA transposons, and 88% attributed to a specific family and superfamily. A large part (74%) of the polymorphisms are due to retrotransposons, consistent with the fact that retrotransposons are the most abundant TEs found in the melon genome (see chapter 1, **Table 1.1**). Interestingly, *gypsy* retrotransposons are more active than what one would expect for their proportion in the genome (46% PM sites versus 36% of annotated TEs). Amongst the DNA transposons, MULE and CACTA families have been the most active, with MULEs being more highly represented than their relative proportion of annotated TEs (12.78% vs 9.6%) would suggest, and CACTAs less represented than expected (3.84% vs 8.12%)

(see **Table 1.1** for percentages of annotated transposons).

A total of 33 families of the 323 families annotated in melon are responsible for 80.9 of the polymorphic sites. (**Table 2.5**). These are mainly *gypsy*, *copia* and MULE type TEs, with the two most abundant being *gypsies*.

family	number of PM sites
CM_gypsy_106	227
CM_gypsy_126	173
CM_MULE_10	133
CM_gypsy_116	99
CM_MULE_13	86
CM_gypsy_97	60
CM_copia_96	58
CM_copia_0	54
CM_PIF_8	54
CM_copia_45	46
CM_copia_28	43
CM_copia_70	41
CM_copia_83	35
CM_gypsy_137	35
CM_gypsy_88	32
CM_copia_44	29
CM_CACTA_12	24
CM_gypsy_58	20
CM_copia_85	18
CM_gypsy_0	16
CM_gypsy_110	15
CM_copia_35	14
CM_gypsy_103	14
CM_gypsy_77	14
CM_gypsy_10	13
CM_gypsy_118	12
CM_gypsy_43	12
CM_copia_15	11
CM_gypsy_31	11
CM_CACTA_15	10
CM_gypsy_28	10
CM_gypsy_9	10
CM_MULE_8	10

Table 2.5: families responsible for the majority of PM sites

In order to get an idea to which degree these polymorphisms are shared between lines, we counted the number of sites that present a TE in exactly one, two, three, four, five or six lines (**Table 2.6)**. It is clear that most polymorphic TEs are unique to a particular line. This is consistent with our previous conclusions that the majority of TEs in melon are recent, as they are more recent even than the divergence of these varieties.

	number of lines sharing the TE insertion					
	1	2	3	4	5	6
retrotransposon	1259	212	103	68	88	301
DNA transposon	360	66	26	42	25	85
uncategorized	52	20	13	5	2	4
total	1671	298	142	115	115	390

Table 2.6: count of lines sharing a PM site

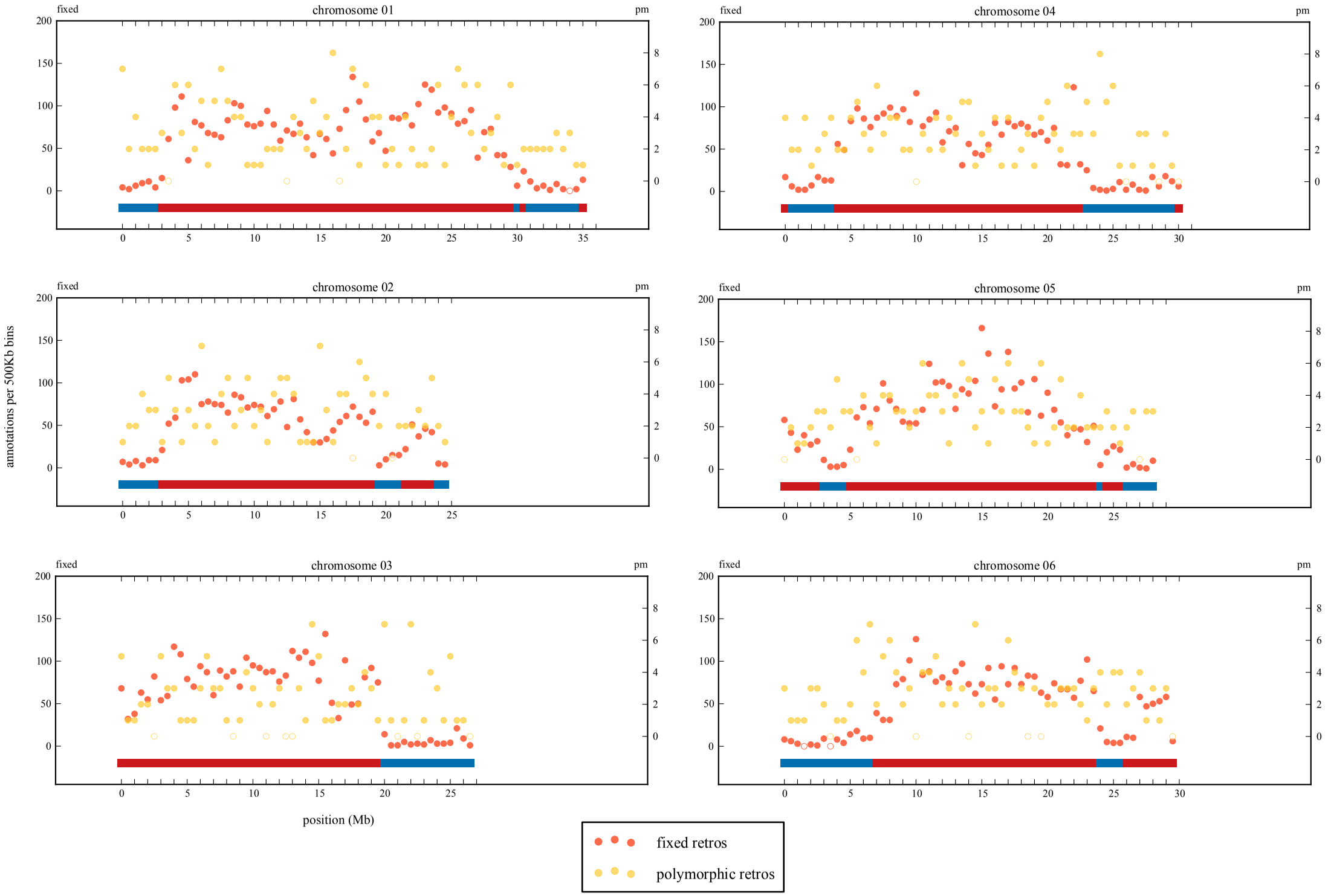
Distribution of TE polymorphic sites along chromosomes

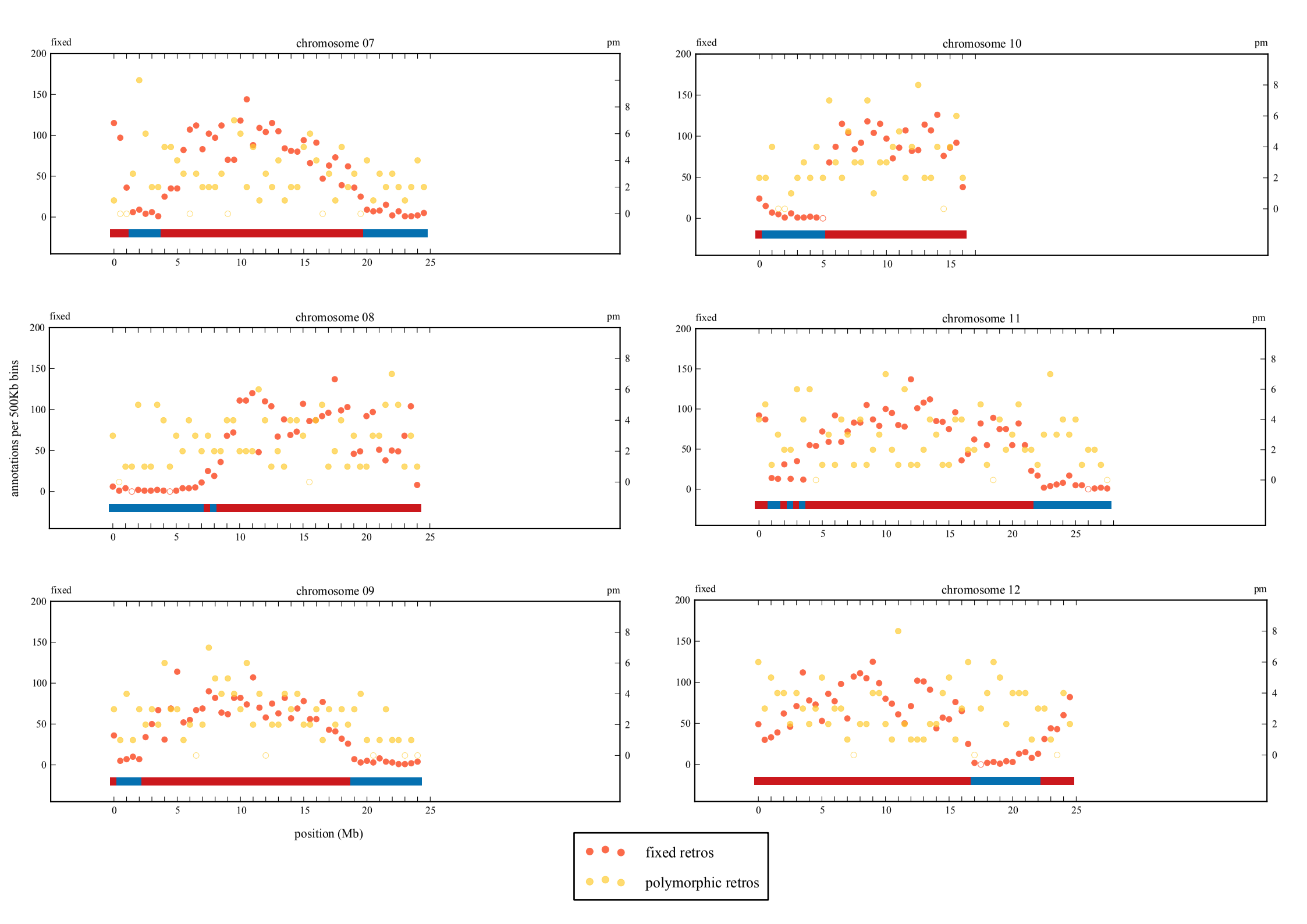
The fixed insertions are those that are older, and have not been selected against, while the polymorphic ones are more recent. In this sense the distribution of polymorphic sites shows us what recent activity has been, and that have had little time to be eliminated. Any differences in chromosomal distribution between old and recent TEs will reveal the selection pressures they are subjected to.

In order to explore whether the polymorphic (PM) TE sites follow the same distribution as the fixed ones, I plotted the frequency of fixed and PM sites in gene-rich and TE-rich regions of the genome, as defined in Chap 1 (see Figure 1.7). I chose to analyze only retrotransposons as DNA transposons are found in too few numbers for statistical analyses.

next two pages:

Fig 2.8 chromosomal distribution of retrotransposon polymorphic sites


chromosome 01
chromosome 04
chromosome 02
chromosome 05
chromosome 03
chromosome 06
fixed
pm
annotations per 500Kb bins
position (Mb)
fixed retros
polymorphic retros


chromosome 07
chromosome 10
chromosome 08
chromosome 11
chromosome 09
chromosome 12
fixed
pm
annotations per 500Kb bins
position (Mb)
fixed retros
polymorphic retros

The chromosomal distribution plots show that the fixed retrotransposons correlate with pericentromeric regions (red), while the PM sites are more scattered. In order to determine whether this pattern is maintained for both *copias* and *gypsies*, I calculated the frequency with which they occur in either type of region **(Table 2.8)** The fixed sites are clearly correlated with TE-dense regions for both *copia* and *gypsy* type retrotransposons (p = 0). Polymorphic *copia* elements do not show a statistically significant association to either type of region, telling us that the drastic difference in fixed TE accumulation in pericentromeric regions is mostly due to selective pressures. The polymorphic *gypsy*-related sites show a slight but statistically significant increase in TE-rich regions, which is consistent with the fact that they tend to target heterochromatic regions for their insertion. In both cases, the differential distribution between fixed and recent TEs shows that TEs are eliminated preferentially from gene-rich regions.

		frequency per bin		fold difference	p-value
		gene-rich	transposon-rich		
COPIA	fixed TEs	2.30	27.21	11.83	0
	polymorphic TEs	0.94	0.98	1.04	0.35
GYPSY	fixed TEs	3.64	45.51	12.50	0
	polymorphic TEs	1.57	1.90	1.21	0.0047

Table 2.7: association of fixed and PM TEs to pericentromeric regions

Locating insertions and deletions on the phylogenetic tree

The phylogenetic relationships of the seven varieties has been established using SNP data (Walter Sanseverino, unpublished) and we have taken advantage of these to establish when the TE movements have occurred that lead to the polymorphisms we observe today. For this analysis we did not consider DNA transposon related polymorphisms as they can both insert and excise, and differentiating absence due to excision and absence due to lack of insertion requires careful inspection of the empty site for

possible traces of TSDs. While retrotransposons can also be deleted (as we have seen in Chap1), it is safe to assume that in the timeframe considered – evolution within a given species – it is more probable that an absence is due to lack of insertion at that site. Thus, given the combinations of lines which show presence or absence of a TE at a given PM site, we placed them as insertions in the phylogenetic tree if this was most parsimonious solution. We assumed insertions to be more likely than deletions, but for those PM sites that could not be explained as a single insertion event in a given branch of the tree, we considered a deletion more likely than multiple independent insertions at the same locus. We were able to place 84.3% of the polymorphic sites on the phylogenetic tree and more than half (55%) of the non species specific sites (**Figure 2.9**). This yields a history of insertions and deletions throughout the evolution of these varieties.

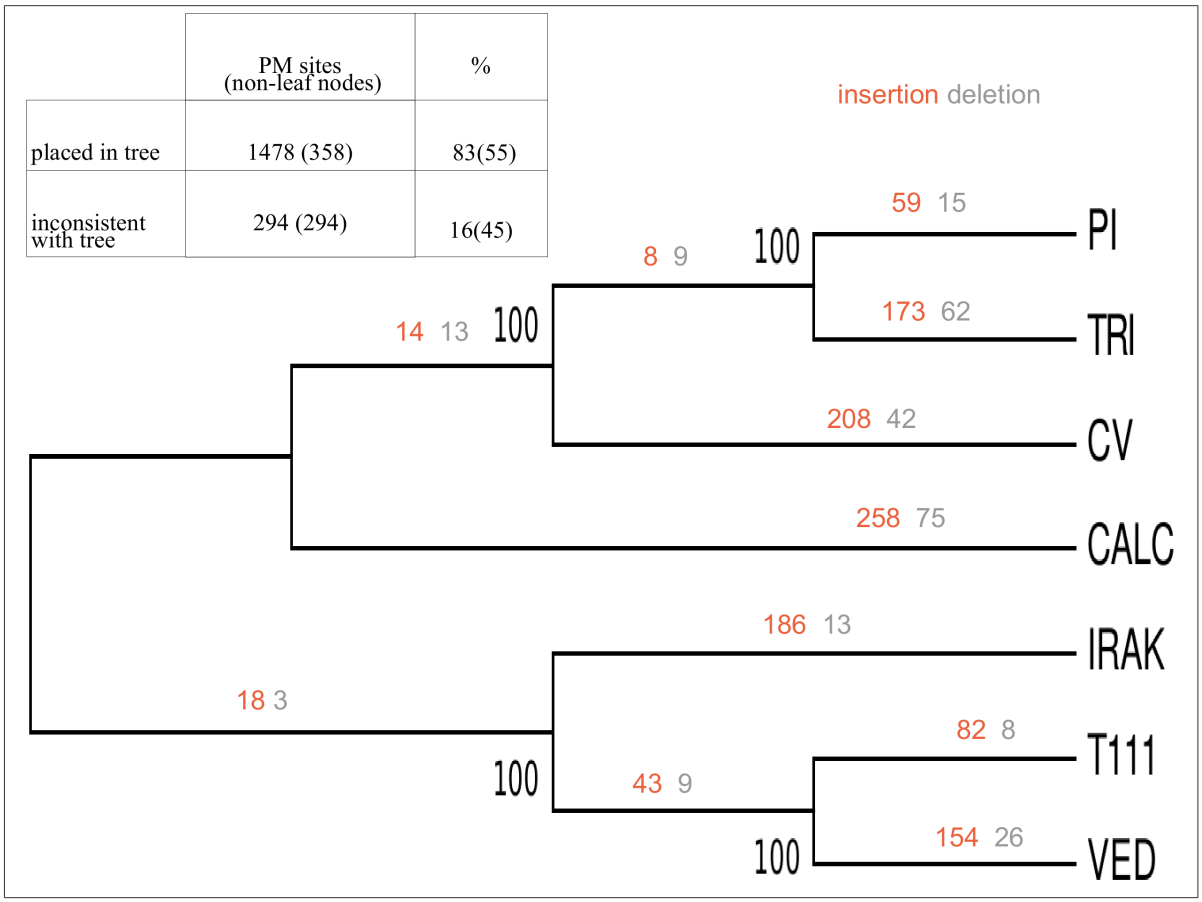


Figure 2.9: retrotransposon PM sites placed on phylogenetic tree

In all the branches of the phylogenetic tree, we can see that there has been more insertion events

than deletions, though their number varies greatly between lines: Calcutta having the most insertions (258) and PI the least (59). Since we do not know the length of the branches in this tree, however, we do not know whether these differences are due to higher rate of activity or just more time past since divergence with the closest variety. In order to evaluate rate of transposon expansion, we can make the assumption that rate of deletion is constant. Indeed, while rates of LTR retrotransposon deletions varies from one species to another (see (H. Wang and Liu 2008) for a comparison of LTR removal rate in Medicago and rice) these rates seem to be functions of various genome-specific features, such as recombination rates and properties of the LTR retrotransposon families themselves. Thus I believe it is safe to say that the removal rates would be similar between different varieties of the same species. In this manner, the rate of TE expansion can be calculated as the ratio of insertions to deletions within a given lineage **(Table 2.8).** Across the seven lines considered, there is a range in the ratio of inserted retrotransposons to deleted retrotransposons, showing us that different varieties are witnessing different degrees of recent TE activity. Even more clear is the difference in insertion rate for the older nodes corresponding to the *agrestis* and *melo* sub-species. Indeed, the insertion rates are on average 5.5X higher in *melo* than in *agrestis*, and T111 and Vedrantais by far share the highest number of insertions.

	num insertions	num deletions	ratio of insertion / deletion
cv	173	62	2.790
tri	258	75	3.440
PI	59	15	3.933
calc	208	42	4.952
ved	154	26	5.923
t111	82	8	10.250
irak	186	13	14.308
PI-TRI	8	9	0.889
PI-TRI-CV	14	13	1.077
T111-VED	43	9	4.778
T111-VED-IRK	18	3	6.000

Table 2.8: insertion rate in different melon lineages

Varieties of subspecies melo and agrestis are shown in blue and red, respectively

Genes affected by TE polymorphisms

Of the 2505 polymorphic sites identified in the 12 annotated pseudochromosomes (226 of the 2731 total PM sites occur in unanchored contigs) across the 7 lines, 26% of these are found within genes **(Table 2.9)**. Insertions in exons account for 62.4 % of the PM sites in genes, implying a potential mutagenesis role for these polymorphisms. This information is a great resource for studying the impact of transposition on phenotypic variation, especially with respect to cultivation and domestication. Of particular interest are the polymorphisms between T111 and VED, two cultivated varieties that are very similar genetically yet with very different phenotypes.

	total PM sites	PM sites in genes	% total	PM sites in exons	% total
all 7 lines	2505	656	26.19	401	16.01
T111 / VED	683	165	24.16	112	16.40

Table 2.9: TE PM in genes

2.4 Discussion

Accurate tool to detect TE insertions

We have developed a software tool to detect TEs present in a sequenced sample that are absent in the reference genome. Both *in silico* and molecular verification has shown that these predictions are highly reliable. The software itself is easy to use and the filtering criteria established can be extrapolated to other genomes and sequencing libraries, making it a useful tool for studying TE polymorphisms in many other contexts. The fact that this software uses only the mapped reads and a TE annotation, without any need of a set of consensus TE sequences or a re-mapping step, means that it is particularly well suited to be integrated in large-scale SV detection pipelines. Analysis of the undetected simulated insertions shows that these are mostly due to lack of mapped reads at the simulated insertion site. This usually is due to the lack of mapped reads in this area, meaning this sequence was either un-sequencable or un-mappable. In general, these areas correspond to low-complexity and/or A/T rich regions which have been called the "genomic dark matter" (H. Lee and Schatz 2012).

Retrotransposons have been largely active in recent evolution

Retrotransposons, and especially gypsy elements, are largely responsible for the recent TE activity in the melon genome. About 10% of the annotated families are responsible for 80.9% of the polymorphisms observed, meaning that the majority of the annotated families probably do not contain any active elements. Interestingly, the most active retrotransposon families all have elements with 2LTRs that have been dated as having inserted within 1MYA, less than the average of the LTR retro population in this genome. This indicates that it is a subset of the LTR retro population that is active recently, and different elements than those that were active previously. Also, the most active retrotransposons (responsible for > 50 polymorphic sites) all have fewer than 8 full copies within the genome. This is not a causal relationship since it is not the case that all low copy-number families are more active, and there are many factors that come into play as to the activation of a transposon. However silencing mechanisms such as small RNAs are often a function of copy number, and perhaps one factor is that these families have not reached the critical copy number at which they are silenced.

Differential chromosomal distribution of fixed and polymorphic sites reveal selection pressures operating on TEs

The difference we observe in the chromosomal distribution of fixed and polymorphic retrotransposons is a great insight into the process of selection against TE insertions near genes. With just the static image of TEs in a genome, we can see that these are anti-correlated with genes, but this is the result of two factors: insertion preference and selection. The fact that newer TEs are not concentrated in heterochromatic regions allows us to see that *copia* retrotransposons, as well as *gypsies* to a lesser extent, insert in a much different distribution. This shows us that the major force shaping the chromosomal distribution of TEs is actually selection, rather than insertion bias. Besides insertional mutagenisis – that one would rarely be able to observe because so deleterious – transposons, by virtue of being targets of silencing mechanisms, have been shown to affect the chromatic state of nearby genes (Ahmed et al. 2011). It has been proposed that this effect is that which is selected against, and in *Arabidopsis* is has been shown genome-wide that methylated TEs near genes are under stronger purifying selection than unmethylated ones, and selection agaisnt them leads to the differential distribution between recent and old TEs (Hollister and Gaut 2009). To our knowledge this is the first analysis that uses polymorphic TEs to investigate differential distribution of recent and old transposons, and thus revealing at a smaller scale the timeline of selection that leads to the final distributions observed.

Different varieties have witnessed varying degrees of retrotransposon activity

Analysis of the polymorphic sites with respect to the phylogenetic tree has allowed us to place in time over half of the insertions and deletions of retrotransposons, and in this manner derive an insertion rate for each branch of the tree. We observe a higher insertion rate over the *agrestis* clade compared to the *melo*, which is interesting as the T111 and VED varieties are the two most cultivated ones. In addition, they are the most similar on a gene level yet vary greatly phenotypically. Therefore an investigation into their patterns of polymorphism and phenotype offers an exciting opportunity to investigate the relationship of TE activity and phenotype on a genome-wide level. About a fourth of the TE PM sites are within genes, and of those 62% within exons. This percentage holds true for polymorphisms between T111 and VED, and would be a good start to investigating the impact of TEs

on gene evolution and domestication. The fact that T111 and VED share so many common polymorphisms either indicates especially high activity in their progenitor variety or can also be a sign of introgressed sequence. Indeed, these varieties are not reproductively isolated and their history of domestication might included crosses. Preliminary SNP data indicates that there are certain chromosomal regions very similar between these two varieties (Walter Sanseverino, unpublished) and these particular regions would be those introgressed between the two species. The high degree of DNA similarity yet difference in phenotypes of these two cultivated varieties makes them ideal candidates to asses the importance of TE movement in evolution.

CHAPTER 3: IMPACT OF TRANSPOSITION

3.1 Introduction

As outlined in the introduction we have seen that besides the most obvious impact of insertional mutagenesis, TEs can have a range of effects, from chromatin organization to chromosome structure to changes in gene expression. The last of these is very interesting in terms of evolution because changes in expression patterns are less likely to be lethal or quickly eliminated, and many examples of adaptation come from changes in temporal or spatial expression patterns. Transposon activity can affect gene regulation in a range of more or less subtle ways. As TEs are targets of silencing they can modulate gene expression through transcriptional gene silencing by modifying the chromatin state of nearby genes. An interesting example is the silencing of the FLC gene by small RNAs directed to a transposon insertion in an intron, which confers vernalization-independant flowering to lines containing this allele (Liu et al. 2004). This phenomenon has been analyzed genome-wide in Arabidopsis (X. Wang, Weigel, and Smith 2013) showing that genes near siRNA-targeted TEs have lower expression levels on average. Some TEs such as retrotransposons carry their own promoters which can drive the expression of nearby genes, causing new inducibility properties, or spatial or temporal expression patterns. For example, blood oranges owe their color to the insertion of a LTR retrotransposon which drives the expression of a MYB transcription factor (TF), itself inducing anthocyanin production (Butelli et al. 2012). An insertion of a TE downstream of a gene can generate antisense transcripts and cause it to be targeted for post-transcriptional degradation, and readthrough transcription into nearby genic regions can also modify their expression (Hernández-Pinzón et al. 2009).

Intrinsic qualities of the TE itself can make it likely to be co-opted into being an expression regulator. This is true regarding both the DNA sequence of the element as well as the properties of the proteins it encodes. Transposases contain DNA binding domains and these have been often exapted for different cellular functions. For example, the DNA binding domain of a MULE transposase has been domesticated into the genes FHY3 and FAR1 in *Arabidopsis thaliana*, two transcription factors which activate several genes involved in far-red light and circadian signaling (Hudson, Lisch, and Quail 2003).

Transposons contain their own promoters which can drive the expression of nearby genes, and their encoded proteins can even act as TF themselves. However, their contribution to the transcriptional regulation of the genomes they inhabit is not restricted to these effects. TEs have been shown to be associated with the binding sites of several master transcription factor binding sites (TFBS), such as ESR1, TP53, MYC, RELA, POU5F1, SOX2, and CTCF in humans (Bourque et al. 2008). In this study they identify by ChIP-seq the regions throughout the genome that are bound by each of these transcription factors, and find that a significant portion of these overlap with annotated repeats. Most interestingly, some regions pulled down by several of these TF individually are associated with the same mobile element, meaning it contains several TFBS. These given TEs are better "progenitors" of the TFBS sequences in question than from any other promoter sequence in that these sequences can arise through less mutations from the consensus of the TE. Taking these data together it can be postulated that TEs have the capacity to generate the combinatorial patterns of TFBS necessary for complex regulation of gene expression. In a similar way it has been shown that the binding of the Estrogen Receptor alpha (ER) is enriched in TEs and that these also host combinations of sites corresponding to transcriptional regulators known to interact with ER (Testori et al. 2012). TE insertions carrying the TFBS and that are in the vicinity of conserved estrogen-regulated genes are themselves conserved in mouse and human, suggesting that those genes are regulated through the TFBS located within TEs (Testori et al, 2012). On the other hand, TEs have supplied BS for OCT4 and NANOG which are not conserved between human and mouse, suggesting that these TEs have wired new genes into these regulatory networks in a lineage-specific manner (Kunarso et al. 2010). In some cases, the same TFBS may be present in different families of TEs. This is what has been seen for the p53 transcription factor, whose binding sites are associated with two families of retrotransposons in hominids (Wang et al. 2007). Given the phylogenetic relationship of these families, the authors postulate that an ancestral family acquired the p53 binding site before the two families diverged. The insertions of p53BS – carrying elements are further from genes than one would expect by chance, suggesting insertions that influence gene expression have been selected against. However, in a handful of cases, transposons have supplied the p53 binding site to p53-regulated genes. The p53 BS has also been shown to be present in a third family type, Alu repeats, which also contribute p53 binding sites to some p53-regulated elements, In this case there is evidence that these site have arisen through mutation in the Alu sequence, rather than being present in a progenitor sequence (Cui, Sirotin, and Zhurkin

2011). Interestingly, some zebrafish orthologs of these genes are also driven by a transposon-carried p53 TFBS though an entirely different, lineage-specific, transposon (Micale et al. 2012). The fact that different TEs containing this TFBS have contributed it to the regulation of the same genes by independent insertion events has been explained as an example of convergent evolution (Micale et al, 2012).

These examples of TEs contributing regulatory sequences and thus influencing gene expression point to their capacity as a powerful driver of evolution, and transposon-mediated rewiring of gene expression networks is likely to have contributed directly to the evolution and diversification of mammals (Gifford, Pfaff, and Macfarlan 2013). The replicative nature of transposons, and their (mostly) random insertion patterns, make them a particularly apt vehicle for shuffling regulatory sequences and relocating binding sites. These exapted functions, and their recurrence, might be an explanation for why a significant fraction of the conserved (and therefore functional) non-coding sequences in the eutherian clade are TE-related (Mikkelsen et al. 2007; Lowe and Haussler 2012).

For these reasons, our curiosity was piqued when we happened across a MITE in *Arabidopsis thaliana* which contains the binding site for a TF. An external collaborator to our lab, Jordi Payet, designed a MITE-finding tool (SUBOTIR, unpublished) which we used to annotate MITEs in melon. He tested the tool using the well-annotated *Arabidopsis thaliana* genome, and while analyzing the results he noticed that several of the sequences identified contained a particular short repeated sequence: TTTCCCGCCAAA. Further analysis revealed that this sequence fits the different consensuses proposed for the E2F binding site (E2F BS) NTTssCGssAAN (Vandepoele et al. 2005), NTTCCCGC (Naouar et al. 2008) and TTssCGss (Ramirez-Parra, Fründt, and Gutierrez 2003). Given that the E2F binding site regulates such crucial functions as DNA replication and cell cycle (De Veylder et al. 2002; Ramirez-Parra et al. 2004) and that MITEs are known to be able to rapidly amplify and generate very high copy-number families, and also have a propensity to insert near genes, we decided to to analyze these MITEs in more detail. The association of a TF BS to TE sequences such as the one we describe has not yet, to my knowledge, been reported in plants, and in this third chapter I'd like to describe our investigation of this phenomenon. This project has been a rich collaboration with other members of my lab as well as external collaborators, and so when it is necessary to understand the

whole picture I will mention results obtained by others, pointing it out clearly when it is the case. For this project I did not develop any novel software so all the related methods can be found in the Materials and Methods section.

3.2 Results

The E2F binding site is found at a high frequency in transposons in Arabidopsis thaliana

The sequence TTTCCCGCCAAA, which fits the consensus NTTssCGssAAN for the E2F binding site, was found in several copies of MITE elements in the *A. thaliana genome*. In order to determine if the number of these sequences in TEs make up a significant part of the total number of E2F BS present in the genome we compared the total number of instances of this sequence to those present in transposon sequences. Surprisingly, 89% of these sequences are found in transposable elements, while these only comprise 20% of the genome (www.arabidopsis.org). In order to determine whether this was the case for the 9 other sequences fitting the consensus NTTssCGssAAN, we also calculated their percentage found in TEs (Table 1). We can see that it is this specific sequence, and not others of the consensus, that is most concentrated in TEs, suggesting that this particular sequence has been specifically amplified in *Arabidopsis* within TEs.

sequence	*A. thaliana*		
	count not TE	count in TE	%in TE
TTCCCGCCAA	189	1668	89.8
TTCCCGCGAA	63	26	29.2
TTCCCGGCAA	99	19	16.1
TTCCCGGGAA	99	17	14.7
TTCGCGCCAA	44	30	40.5
TTCGCGCGAA	6	2	25.0
TTCGCGGCAA	48	4	7.7
TTGCCGCCAA	84	45	34.9
TTGCCGGCAA	28	5	15.2
TTGGCGCCAA	36	56	60.9

Table 3.1: frequency of E2F consensus sequences in and out of TEs in A. thaliana

In order to analyze if this phenomenon is specific to *A thaliana* we performed the same analysis in four related *Brassica* species: *Capsella rubella*, *Brassica rapa*, *Thelungella halophila* and *Arabidopsis lyrata*.

	TTCCCGCCAA			
	count not TE	count in TE	% in TE	% TEs in genome
A. thaliana	189	1668	89.8	20.30
A. lyrata	365	6063	94.3	23.33
C. rubella	209	2145	91.1	11.74
B. rapa	350	2579	88.1	11.71
T. halophila	260	129	33.2	28.48

Table 3.2: frequency of sequence TTTCCCGCCAAA in and out of TEs in five related Brassica genomes

This revealed that the prevalence of the sequence TTTCCCGCCAAA outside TEs is relatively constant in the five genomes but in all of them except *T halophila* a much higher number of these sequences is found within TEs. (**Table 3.2**)

Thus far we know that there is one sequence of the 10 that fit the consensus binding site of the E2F TF that is found at much higher frequency than expected in mobile elements in four related *Brassica* genomes. Here we cannot help but notice that in the genomes that show a concentration of the E2F sequence in TEs also show a considerably higher number of total sites. Due to the replicative nature of transposons, we wondered whether these sequences might have been amplified by transposons as well.

The sequence TTCCCGCCAA and not others of the E2F consensus has been amplifie by transposonsd in four brassica genomes

To account for differences in genome size, and taking *Oryza sativa* as an outgroup for comparison, we plotted the frequency per megabase of each of the sequences of the E2F consensus in these six genomes (**Figure 3.1**). We found that compared to rice, only the sequence TTCCCGCCAA and not others was amplified in the four *Brassica* genomes except for *T. halophila.*

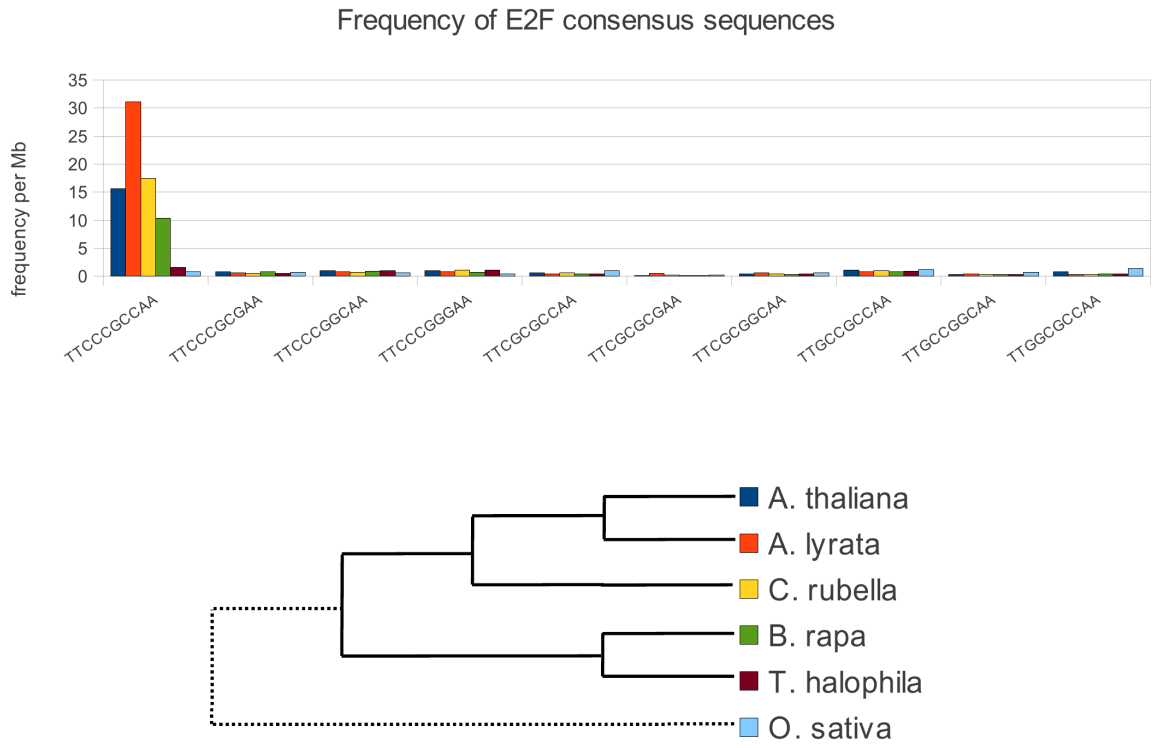


Figure 3.1: frequency per megabase of the 10 sequences fitting the E2F consensus TTssCGssAA

This analysis shows that the sequence TTCCCGCCAA has been amplified by transposons in four *Brassica* genomes. In order to clarify the relationship between the amplification of the sequence TTTCCCGCCAAA and its concentration in transposons, we plotted the relative frequency of the set of sequences up to 2 mutations away from TTCCCGCCAA both within and out of transposons. **(Figure 3.2)**. For each of these genomes, there has clearly been an amplification of the sequence TTCCCGCCAA and not any of the other similar sequences. In each of the cases, the pattern of amplification genome-wide is maintained when restricted to the TE sequences, and disappears when looking at non-TE sequences. This analysis shows that while the sequences found outside TEs are present in relatively similar frequencies, the sequence TTCCCGCCAA is the only sequence present at higher frequency in the four genomes where it has been amplified. This shows that this sequence has been amplified specifically by transposable elements in all four genomes and that it has been strongly conserved as we do not see signs of mutated sequences accumulating after amplification.

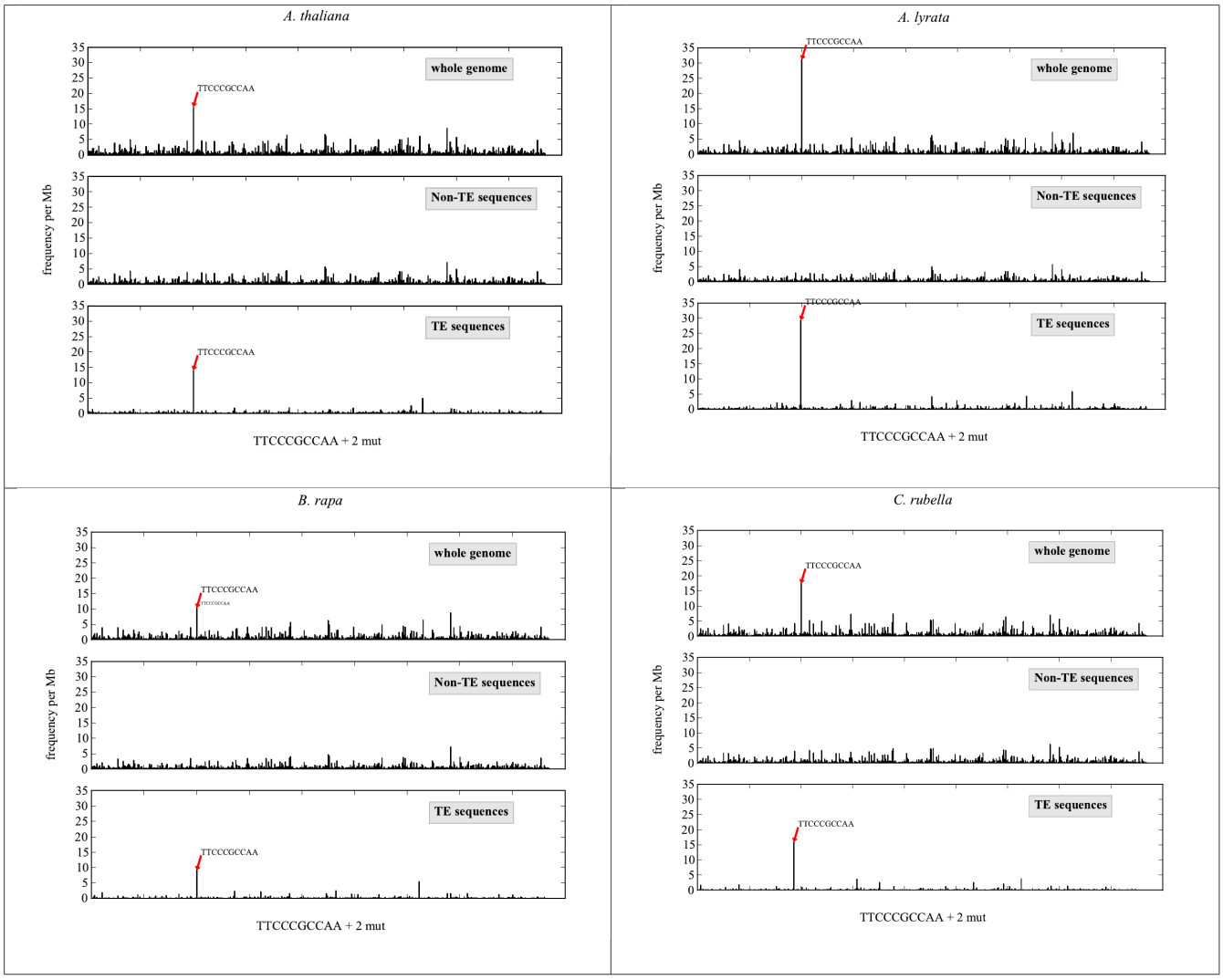


Figure 3.2: frequency per Mb of all sequences up to 2 mutations away from TTCCCGCCAA

From these results we conclude that one of the 10 sequences that fit the E2F consensus has been amplified by transposable elements in four related genomes: *A thaliana*, *A lyrata*, *B rapa* and *C rubella*, but not in the closely related species *T halophila* or the outgroup *O sativa*.

These observations lead to the following questions: how has this amplification arisen? What are the possible impacts of the association of such a crucial TFBS with TEs?

The E2F sequence has been amplified by different families in different genomes

In order to gain insight into how and when the E2F sequence was captured and amplified by TEs, we performed a comparative analysis of the main TE families which contain it in each of the 5 Brassica genomes. For this, a crucial information is the family definitions of the elements within each genome. However, TE annotations in these genomes – besides *A thaliana* – have not been well-curated, and are mainly based on masking the genome with RepeatMasker (Hu et al. 2011 for annotation of *A lyrata*, Materials and Methods for annotation of *R rapa*, *C rubella* and *T halophila*). The family names attributed to the different repeats are based on sequence similarities with elements in the Repbase database (http://www.girinst.org/repbase/). Repbase contains well-defined elements but also elements not so well characterized, which makes it risky to rely solely on sequence similarities to define families of elements. We therefore decided to curate the TE annotation and redefine the families of the annotated elements. We clustered all the annotated transposons into families (80% similarity along 80% length) to obtain family "nuclei", then attributed fragments to their respective nuclei when they are at least 60% similar along 20% of their length (using the SILIX clustering software, http://lbbe.univ-lyon1.fr/Overview.html (Miele, Penel, and Duret 2011)), then picked representatives of each family. We then performed pairwise alignments of all the representatives from all the genomes, and thus defined family relationships between the families in different genomes. **Figure 3.3** sums up this data, with each family color-coded and the size of the box proportional to counts in each genome. This analysis was performed by Ankita Chaurasia, a member of our lab, and the following figure summarizes her work.

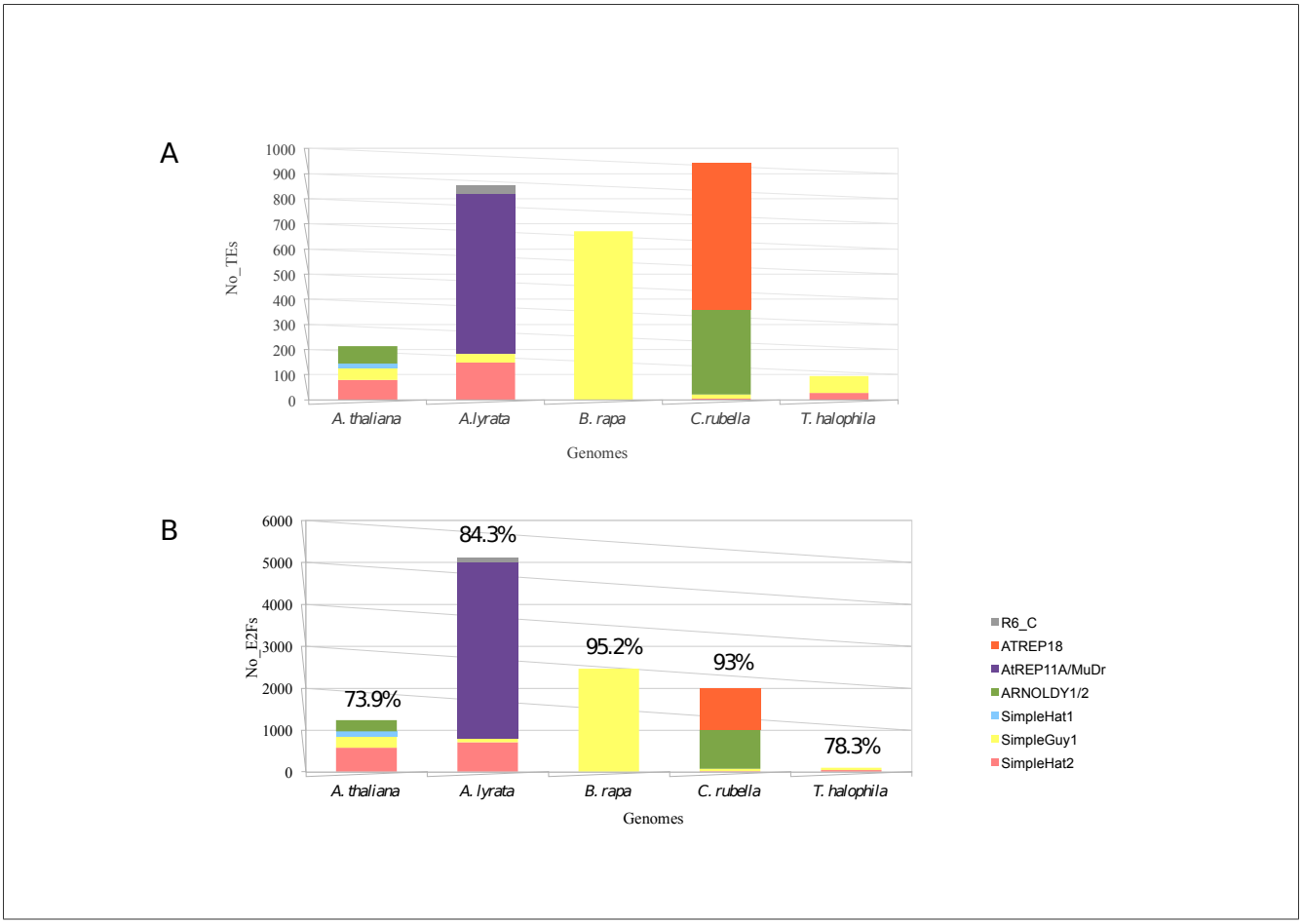


Figure 3.3: families containing E2F sites in related Brassica genomes.

A) Number of copies in TEs families containing E2F sites per genome B) Number of E2F sites in TE families containing E2F in each genome. Over each bar is the percentage of total E2F sites in TEs contained in the families represented in the chart.

In *Arabidopsis* four TE families account for 73.9 % of the total E2F sequence TTCCCGCAA (**Figure 3.3 A**). These four families belong to three different superfamilies of DNA transposons: hAT (Simplehat1 and Simplehat2), PIF/Harbinger (Simpleguy1) and MULE (ARNOLDY1/2). Three of these four families (Simplehat2, Simpleguy1 and Arnold) are also found in other of the species analyzed, while Simpleguy1 is found in all five (**Figure 3.3 B**), suggesting that these families were already present in the ancestral species. However, the different prevalence of each of these families shows that they have amplified to a different degree in each lineage after speciation. Some families,

such as ATREP18 in *C rubella*, are specific to that genome. In total there are seven different TE families, belonging to different DNA transposon superfamilies, that have amplified the E2F motif in the five species analyzed. It is interesting to note here that *T halophila* does have two TE families containing the E2F motif: SIMPLEHAT2 and SIMPLEGUY1, which are also found in other of the genomes analyzed, though these are in too low copy number in *halophila* to have generated a significant amplification of the number of E2F sequences.

Thus the E2F BS is associated with different families of TEs in these five genomes, and these families have distinct histories in each species, leading to drastic amplification of the number of E2F sites in some but not in others. Here we point out that often the number of E2F sites is larger than the copy number of a family: there are therefore several E2F sites in a given transposon. We now take a closer look at the elements in *A thaliana* to elucidate this question.

The E2F sequences have been amplified within tandem repeats

In at least three of the four major TE families characterized in Arabidopsis, the amplified E2F sites are mostly found within a larger 27bp motif repeated in tandem, forming a minisatellite. Examples of sequences from the three main families in *Arabidopsis thaliana* are shown in **Figure 3.4**.

A. SIMPLHAT2

AGGGGTGTCAAAATGGGTCAAAATTCATGGGTCAACTCAACTCAACTCAACC
CATGAACCCTTAAATTTTTTTACTCAAATGAGTTGATGGGTCAAATGGGTTG
ATGAGTCAATTGGTTTGATGAGTAAAATGAGTTGGGTTGTAATGGTTAATGG
TTTCAATGGTTTACCCAATTAACCCATCAAATTTTGTAAATTTGAATTAAAC
CAACTAAAATCTCTAAACCAATGCCAATTTAAGTTTAACAAACACATCTAAA
CCAATTTAATAAAATTAATATTTTTCCAAAATTCTTAAATATACAAGCGATG
CAATT
GAGAAAAAATAAACTCGT
AATTTTTCCACCAAAAAACATAAACCCGT
GATTTTCTCGCCAAAAACGTAAACCCGT
GATTTTCTTGCCAAAAACGTAAATCCAT
AATTTTCCCGCCAAAAACGTAAACCCGT
AATTTTCCCGCCAAAAACGTAAACCCGT
AATTTTCTCACCAAAAACGTAAACCCGT
GATTTTCCCGCCAAAAATGTAACCCCGT
GATTTTCCCACCAAAAATGTAACCCCGT
GATTTTCCCGCCAAAAACGTAAACCCGT
GATTTTCCCGCCAAAAACGTAAACCCGT
GATTTTCCCGCCAAAAACGTAAACCCGA
GATTTTCCCGCCAAAAACGTAAATCCGT
TATTTTTTCGTCCAAAACGTAAACCCAT
GATTTTTCCGTCAAAAACGTAAACCCAT
CATTTTCCGGCCGAAAACGTAAAAACCTAT
AATTTTCCCACCAAAAACGTAAATCCGT
GATTTTCGGTACCAAAAACGTAAACCTGT

GAAAAGTGGAATCTGTAAATATCTTAAGTTTGATGATAATGAATTAATAATG
ATAATTATTTATTATTTTTTTATAATAATAATTAATTAAATTATTACTTAAA
TGGGTTAACCCATTTAACAACTCAACCCATCAAATTAAATGAGTTATGGGTT
GACCCAACTTATTTTGTTAAATAAGTTGGGTCAACCCATAACTCATTTAATT
TTAAACTCATTTGATTATGAGTTGAGTTGAGTTGGGTTACCCATTTTGACACCCCT

B. SIMPLEHAT1

TAGAGATGTTAATGATGGGTTAAAACCCACTGGGTACCCAAAACCCACATAA
AATTTTAAGCCCATAnATCTACTTTTTCTGGAAAAAAAACACAGGTTTAAAA
TGGGCTGGGTACCCAAATAAATAAAACCCATATGGGTTTACCTCATTGGGTT
ATGGGTACCCACGGGTTTTTTTTAAGT

CCTTTTTCAAGTGAATCAA
CATTTTCTTGTCAAAAAAGATTTCA
CATTTTTTTTCTCCAAAACCGCGATTTAA
CAGTTTACCGCCAAAAAAGACATTCAA
CATTTTCCCGCCAAAAAGACATTCAA
CATTTTCCCGCCAAAACGACAATCAA
CATTTTCCCGCCAAAA CGCAATTCAA
CATTTTCGCGCCAAAAACGACATTCAA
CATTTT CCGCCAAAA CGCAATTCAA
CATTTTCCCGCCAAAACGCGATTCAA
CATTTT CCGCCAAAAAT GATTTCA
CATTTTCCCGCCAAAACGCAATTTAA
CATTTTCCCGCCAAAACGACATTCAA
CATTTTCCCGCCAAAAA GATTTCA
CATTTTTCGGTCAAAAACGCAATTTAA
CATTTTCCCGCCAAAATGATATTCAA
CATTTTTCCGCCAAAAACGCAATTTAA
CATTTTTCCGCCAAAAACGACATTCAA
CATTT CCGGCCAAGAAAAGGATTTGA
CATTTTTCCCGCCAAAACGCAATTTAA
CATTTTTCCGTCAAAAAACGCAATTTAA
CAT TTTCCCGCCAAAACGCAATTCAA
CCTTTTTCCCATCAAA

TATATATGTAATACAGAAAATAAATATAATTTAAAATATATTTATAAGCAAA
ATATAGATACAACATACATATAATAGTATATTAACAGTTAAACATAAACTCT
AAGAAAATGATAAAAACCAAGTTGGGTACCCACGGGTAGTCTCGGAACAATC
AGGTTTTTTCGGGCTTTACCCACTAAATCAAAAACCCATCTGGGTTTTCTAA
AGTTGGGTTTTTGGTGGGCGGGTTTATTTTGGGTTCAGGTCGGGCTGGGCTT
TTTTACCCACCACAACATCTCTA

C. SIMPLEGUY1

CTGTTTGTTTCTCCATCCAAATGATCCATCCAAATGAAGATGAAAAATGATG
TTTGTTTTTGACAATTTAACAAACATTTGGATGGATCATCTGGATGATACAT
TTGAATGAGTTTATAAATTTAGGCAAATTTATGAAACTCATTTGGATGAATT
TGATTGAACCATTTGGATGGAGATGGTTCATTCAAAATAACAAAAGATAGAT
TTACCCTCATTTCTAATCATAATTTAGATAATGTTATTTTCTATAAATTTCT
TATTTAATATTTTTAAAATTACATATTGAAACAAAAAAACTCTAAAAATCAT
GAAATCACGGATTATTATTTTATGTAAAAA TCGTAAAATCGTGGAATTGAG

TT TTTTT GGCAAAAACCGCAAAATCATAACATCA
TGTTTTTC GCCAAAACCGCAAAATCA
TGTTTTTCC GCCAAAACCGCAAAATCA
TG TTTTCC GCCAAAACGTAAAATCA
TG TTTTCCCGCCAAAATCGCAAAATCA
TG TTTTCCCGC AAAAACAAAAAATCA
TG TTTTCCCGCCAAAACCGCAAAATCA
TG TTTTCCCA CCCAACCGCAAAATTA
TGTTTTCCCCGCCAAAACTGTAAGATCA
TG TTTCCCCGCCAAAACCATAACATCA
TG TTTTC ACAAATCCGTAAAATTA
TG TTTTCCCGCCAAAACCGCAAACTCA
TG TTTTCCCGCCAAAATCGCAAAATCT
TA TTTCACCGTCAAAACCGTAAAATCG
TG TATTCTCGGCGAAACCGCAAGATCA
TG TTTCCTGCCAAAA CCGTAAAATCA
TG TTTTCCCGCCAAAACCGCAAAATCA
TG TTTTCCCGCCAAAATCGCAAAATCT
TA TTTTACCGCAGAAACCGTAAAATCG
TG TATTCTCGGCGAAACCGGAAAATCA
TG TTTCCTGCCAAAAT CGCAAAATTA
TGTTTTTCCCGCTAAAACCGCAAAATCA
TA TTTTCCCGCCAAAACCATAACATCA
TG TTTTCGC AAATCCGTAAAATCA
TG TTTTCCCGCCAAAACCGAAAACTCA
TG TTTTCCCGCCAAAACCGCAAAATCTT
AA TTTA CCGCCAAACCCGTAAAATCG
TG TATTCTTTGCGAAACCACAAAATCA
TA TTTTCCCGCTAAAACCGTAAAACTG
TA TTTTCCCGCCAATAT GTTTGGTAT
TG ATAATGTGTAATGATGTTGAATATT

TGTCATTTGTCAAAATGAAAGCATGAAGATTTTGGTCATTTTTATCTAAATA
AGAATGTGGATGAATTACAAAATTATCAAACAAATAAATATCTATTTAAATG
CATCATTTAAATTAAATTACAAATAATAAAACAAACATCATTCAAATGTATC
ATTCAAATGAATCATGTAAGTGGATCATCCGAATGGAAATGATAAATGATGA
AACAAACAGGGTAATTTAAATGAAA

D.

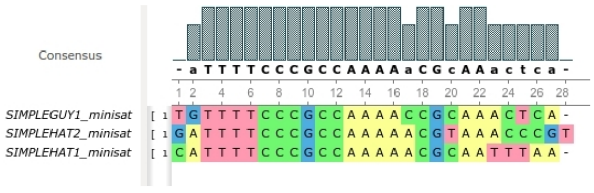


Figure 3.4. *A) B) and C): representative sequences of the three major families of TEs that contain E2F in A thaliana. D) alignment of the minisatellite motifs found in each*

SIMPLEHAT1 and SIMPLEHAT2 are hAT-like elements, while SIMPLEGUY1 is a PIF/Harbinger type transposon, which differ both in their sequence and in the transposase that mobilize them. The minisatellite sequence is more similar between different elements than the rest of the transposon, suggesting they have a common origin.

In order to determine whether the fact that the E2F motifs are in minisatellites holds true for the whole contingent in *A. thaliana*, we identified tandem repeats throughout the *A. thaliana* genome with TRF (http://tandem.bu.edu/trf/trf.html, (Benson 1999)) and thus identified the E2F sequences that are included in a minisatellite. In **Figure 3.5** we represent the whole set of E2F sites in the *Arabidopsis thaliana* genome, and what percentage are within a TE, a minisatellite, both, or neither.

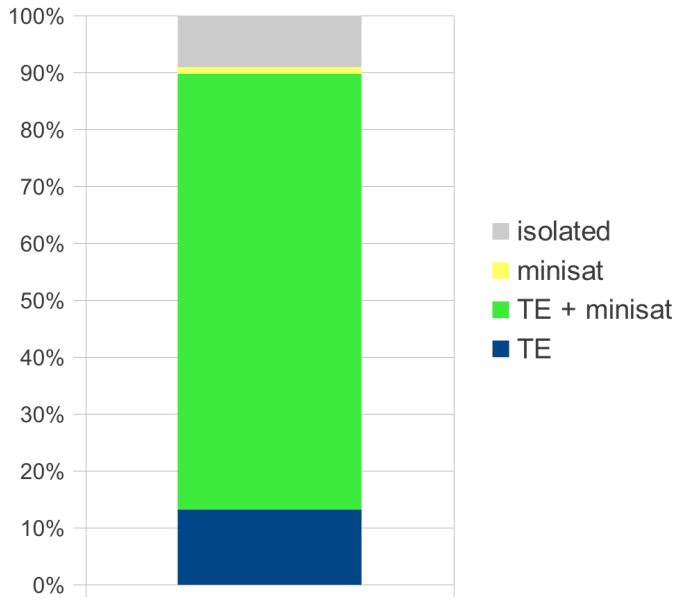


Figure 3.5: Context of E2F sites in A. thaliana

89% of the E2F sites in A thaliana are found in transposons, and of these 88% are found in minisatellites. Three main families: SIMPLEHAT1, SIMPLEHAT2 and SIMPLEGUY1 account for over half (63%) of the E2F sites in TEs in minisatellites. Thus the amplification in *A thaliana* is largely due to these three families.

Taken together we can see that the mechanisms that have led to the increase in number of E2F sites have been twofold: on one hand, these sites are found inside MITE elements, which tend to amplify and generate high copy-number families, and on the other, the sequence is embedded within a minisatellite, itself capable of expanding (and contracting) yielding varying numbers of tandem copies of the motif.

While the amplification of the number of these sites is impressive, it is the fact that the E2F sequence is found in TEs that is the most puzzling. Apart from the sheer number of instances – and the possible impact of sequestering TF protein to these sites – it is the fact that this sequence may be mobilized and relocated in the genome, affecting gene expression, that comes to mind as a possible impact.

In order to address the question of what the impact this association can have, we investigated the relationship of E2F-containing TEs with genes. Since *Arabisopsis thaliana* is the best annotated genome of the five Brassicas studied this part of the analysis is restricted to this species.

Most E2F sites in transposons are far from genes

In order to asses the impact of the amplification of this TFBS, we first investigated whether these TFBS found in TEs might have an impact of gene expression. To do this, we looked at the distance of these sites within TEs with respect to genes.

For a basis of comparison, we plotted the distance of four different TFBS to genes, according to whether it is in the 3' or 5' region of its closest gene (**Figure 3.6**). The occurrences of these TFBS are found overwhelmingly within 100bp of the 5' end of their closest gene.

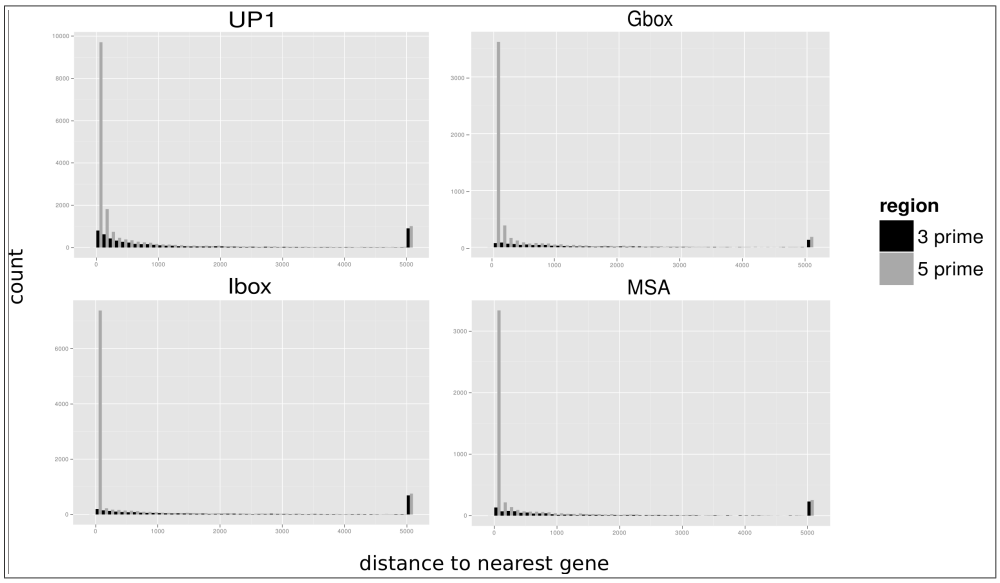


Figure 3.6: distance to closest gene of four TFBS

Similarly, the E2F TFBS outside of TEs are found mostly within 100bp of the 5' region of the nearest gene (**Figure 3.7 A**), suggesting that, in most cases, the E2F transcription factor regulates gene expression by binding very close to the transcription start site, as it seems to be the typical case in the compact genome of *Arabidopsis*. On the other hand, the E2F sites located inside transposons follow a different distribution. Indeed, they are found more evenly distributed between 5' and 3' regions, and clearly farther from genes (**Figure 3.7 B**). This suggests that most E2F TFBS found in TEs will not have a direct impact on the regulation of genes by E2F. However, if this E2F TFBS can be recognized and bound, the resulting sequestration of the transcription factor into non-productive sites would have an indirect effect.

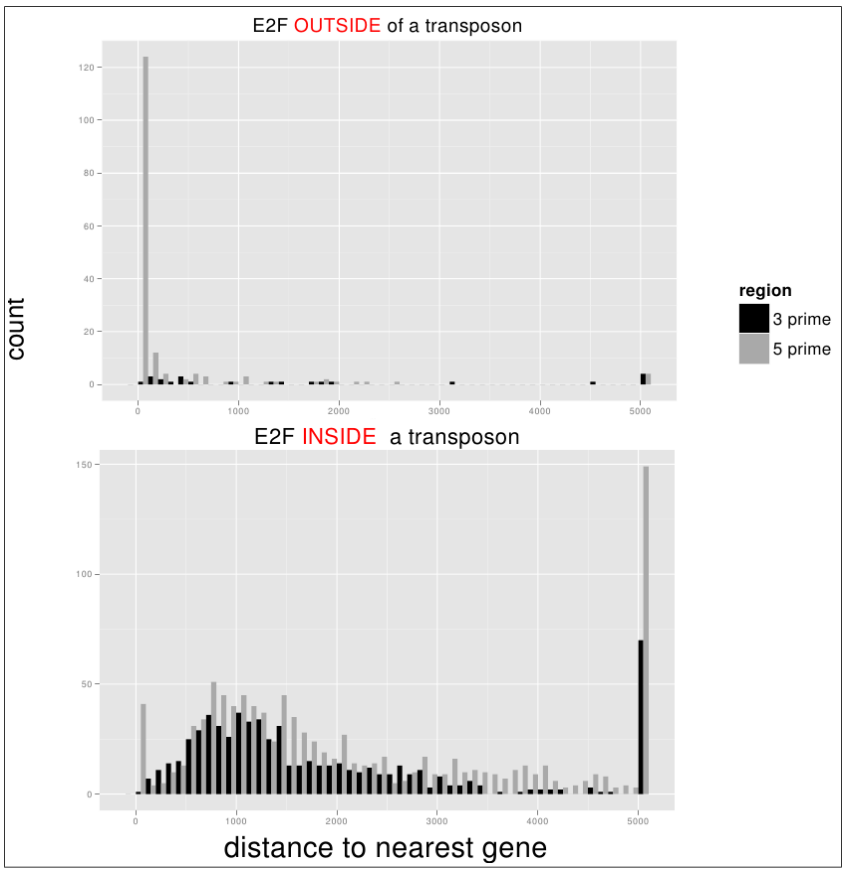


Figure 3.7: distance of E2F sites to genes

While most of the TEs containing E2F are far from genes, there are a few which are close to genes. We decided to take a closer look at their possible role as having integrated a gene into the E2F regulatory network.

Some transposon insertions may have wired new genes into the E2F regulation network

Microarray analyses in plants overexpressing E2Fa-DP identified genes upregulated more than two-fold in these lines with respect to WT (Vandepoele et al. 2005, Naouar et al. 2008). Of these, those that contained an E2F motif in their promoter region were considered as potential E2F targets. These two analyses were performed with different arrays (ATH and tiling), and using different criteria for the promoter region (400 bp and 1000 bp), so we combined the results by taking all genes that had been found to be upregulated in either experiment, and contained an E2F TFBS at less than 1000 bp. Among

the 542 genes that fulfilled these criteria, there are 5 cases where the E2F motif is contributed by a transposon. These may be cases of genes that owe their regulation by E2F to the transposon insertion.

While this *in silico* analysis is an indication that TEs might have contributed an operational E2F TFBS to certain genes, this is wholly dependent of whether the TF protein can actually recognize and bind these sites.

The E2F BS in TEs can be bound by the TF protein

We hypothesized that the "excess" of E2F BS would be inaccessible to the protein TF, avoiding titration of the TF away from its true targets, and that the inaccessible sites would be "hidden" due to their heterochromatic context. We designed a Chromatin Immuno-Precipitation experiment to test this, aimed to determine whether the E2F protein can bind to E2F TFBS in TEs, and if any difference in binding is correlated with different heterochromatic marks. This experiment was carried out by Cristina Vives from our lab.

The results obtained show that, although most of the E2F TSBS located within TEs have heterochromatin epigenetic marks (high levels of H3K27me and low levels of H3K4me2), which differentiate them from the E2F TFBS found outside TEs (which have low levels of H3K27me and high levels of H3K4me2), most E2F TFBS bind the E2F factor irrespectively of them being located within or outside TEs (**Figure 3.8**). In plants over-expressing the E2Fa together with its dimerization partner DPa (both under the control of the CaMV 35S promoter) (De Veylder et al. 2002), all the analyzed E2F TFBS are bound by the E2F transcription factor irrespective of their location and associated epigenetic marks (not shown). These experiments show that all E2F TFBS, including those located within TEs, can potentially bind the E2F transcription factor, and that most of them seem to be occupied *in vivo* in wild type plants and normal conditions. This suggests that transposition of TEs containing E2F TFBS may have a direct impact on the set of genes regulated by this transcription factor. Additionally, any situation in which the binding to the E2F TSFB located within TEs may be increased will have an impact on the quantity of E2F transcription factor that is available to bind the rest of the E2F TFBS, modulating its binding and regulation of E2F-controlled genes.

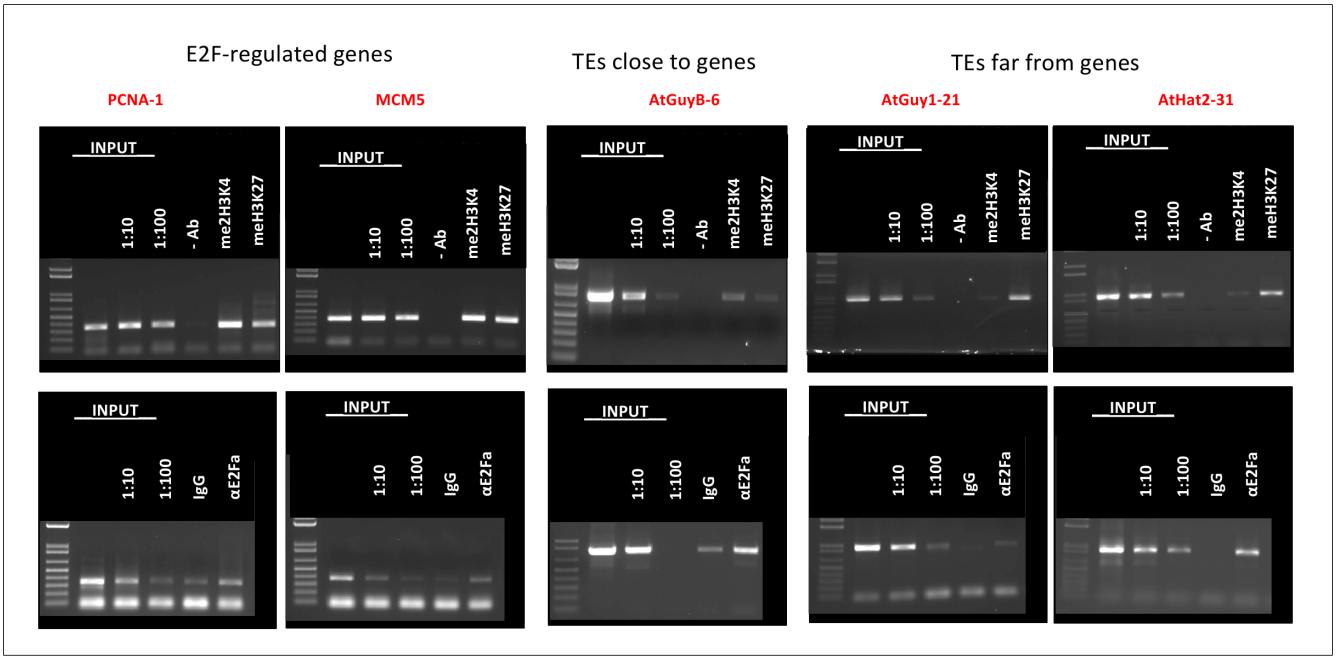


Figure 3.8: ChIP analysis of the epigenetic marks associated to the E2F TFBS (top) and the binding of the E2Fa TFBS to its sites (bottom). Different types of E2F TFBS are analyzed: E2F TFBS located in control E2F-regulated genes, E2F TFBS located within TEs located close to genes and E2F TFBS located within TEs located far from genes.

3.4 Discussion

The E2F BS has been captured by transposons in the Brassica genus and amplified in some genomes

Here we show that the sequence for the E2F binding site is found in transposons in five related genomes: *Arabisopsis thaliana*, *Arabisopsis lyrata*, *Capsella rubella*, *Brassica rapa* and *Thelungela halophila*. It is found in various families, some shared by several (or all) genomes, and others genome-specific. These TEs have amplified the number of occurrences of the E2F BS in all of these genomes except for *T. halophila*. As to the origin of this capture, there are two hypotheses: either it was present in an ancestral family of TEs, which has diverged and led to different TE families, or it has arisen several times independently. The fact that one of these TE families is common to all the genomes analyzed supports shows that at least that family originated prior to the divergence of these species, but

does not enable us to distinguish between a unique or multiple origins. A more in-depth analysis of the TE families in question, and especially determining whether there exist MITEs of the same families that do not contain the repeated E2F motif, would help answer this question.

The amplification mechanisms at work here have been twofold: first, the fact that MITEs themselves can increase in copy number, and also that the E2F motif is found within a minisatellite. As to the origin of the minisatellite, it is important to note that the motif is not identical in the three families is which we have analyzed it, and the most conserved part of the minisatellite is the E2F sequence itself. Here we can imagine two scenarios: either the minisatellite already existed in the ancestral species, and was captured by one (or more) TEs, then evolved within the different TEs. Or, it arose separately in the context of different TEs (independantly of the origin of these TEs). In order to evaluate the support of either of these two hypotheses we need to think a moment about the mechanisms which generate tandem repeats. Minisatellites can be generated by problems in any mechanism that involves new synthesis of DNA, including replication, recombination or repair. In the case of replication, stalling on the lagging strand can cause slippage of the polymerase, which can create a loop when synthesis resumes. Whether the loop forms in the template or the newly synthesized strand causes either a duplication or a loss of the sequence in the loop, respectively (for a review on DNA repeats see Richard, Kerrest, and Dujon 2008). Stalling of the replication mechanism can be due to secondary structure of the single-stranded template DNA, or binding of a protein. The fact that different copies within a same family contain a variable number of minisatellite motifs indicates that this minisatellite has been expanding and / or contracting after expansion of the TEs. Since near identical copies of the motif are made / deleted this maintains the sequence of the minisatellite conserved, and precludes phylogenetic analysis based on the minisatellite sequence.

Since E2F is a protein that is involved in attracting the replication machinery, one can imagine that it could be the cause of replication fork stalling and cause the creation, expansion or contraction of a minisatellite whose motif is centered around the E2F binding site. One could imagine a scenario where various TEs have by chance evolved the E2F BS by random mutations, and then this sequence would become a minisatellite, along with its flanking bp, due to the binding of the E2F protein and stalling the fork. This would explain why the repeated context is different in various elements, and why it is found in very different TEs. Further investigation as to the probability of the E2F sequence arising by chance in these element versus other sequences in the genome would help elucidate this question, as

for example it was shown that the p53TFBS, which is found in different families of TEs, arose by separate sets of mutations (Cui, Sirotin, and Zhurkin 2011).

Impact of E2Fs in TEs on the genome

The impact of the presence of E2F sites in TEs is twofold: one, the impact of relocating these TFBS, and two, the impact of the amplification. As far as the impact on transcription, there are a handful of candidates: genes that are putative E2F targets and who's closest E2F BS is contributed by a TE. The fact that the TF can bind to sites in TEs *in vivo* confirms this possibility. This will not be the first time that a TE wires a new gene into a regulatory network, and it has actually been postulated that its the crucial TF like OCT4 or NANOG that need this combinatorial and randomization function, and that is what has permitted them to become master TF. Further investigation into whether these insertions are conserved in other genomes, or other varieties of *Arabisopsis thaliana*, will provide insight into the functionality of these insertions. Perhaps even the fact that it is in a minisatellite, which confers variability, gives it more potential for evolvability (Vinces et al. 2009).

Another potential effect is the titration of the protein away from its “real” targets. We first hypothesized that the E2F sites in TEs would not be accessible, hidden away in heterochromatin. Surprisingly, even though E2FBS in TEs do show heterochromatic epigenetic marks, they can be bound by the TF. However, this might not be the case in all tissues of in all stages of the cell cycle, and makes us wonder whether the genome has another way of regulating its binding.

Impact of E2FBS on TEs

The next question is what is the potential impact of the presence of the E2F site on the TEs that contain it. MITEs amplify to large copy numbers of almost identical copies through an unknown mechanism. Though many MITE families do not contain the E2F it is a possibility that those that do have been aided by it: the E2F protein can attract the endoreduplication machinery, and transposition in an endoreduplicated state could lead to amplification. The fact that the same TEs with E2F have amplified or not in different genomes means that any advantage the E2F site may confer is specific to a particular genome.

General Discussion

GENERAL DISCUSSION

In this study we have approached the question of deciphering the evolution of transposable elements and their impact the host genome from various angles. On one hand, we have investigated the transposon landscape and assessed its dynamic features in a newly sequenced crop plant. This yielded insight into the dynamics of transposon evolution in this genome, and the selective forces that have shaped the transposon landscape. On the other hand, we have delved into a particular case in the model plant Arabidopsis of a master transcription factor that has been captured and amplified by transposons. This study gives an example of one of the many manners in which TEs can impact gene regulation.

In the melon genome, we have determined that the current transposon content are elements that were active recently. Notably, most elements are specific to the melon lineage, and have expanded after the split with its close relative cucumber. Phylogenetic analysis shows that DNA transposons are not only more abundant in melon but most families are melon-specificThe dating of the LTR retrotransposon insertions is consistent with a recent expansion of these elements. Indeed, all characterized retrotransposons inserted after the melon-cucumber split, and are likely partially responsible for the difference in genome size between these two species. Interestingly it seems that the highest LTR retrotransposon activity happened around 2 MYA and that there has been less activity in the very recent evolution of this genome, unlike other plant genomes such as *Medicago*, rice or *Arabidopsis*. However the history of each family is distinct, implying that there are features of the individual families that influence expansion and removal rates, even though the general trends point to recent activity. These results suggest that transposable elements have played a major role in shaping the melon genome in recent evolution. .

In order to get a dynamic view of the impact of TEs in recent genome evolution, we have exploited available resequencing data of seven melon varieties. In order to perform this comparative genome analysis, we have developed a software tool to detect TEs present in a sequenced sample that are absent in the reference genome. Both *in silico* and molecular verification has shown that these predictions are highly reliable. The map of polymorphisms reveals that a small fraction of the families annotated are actually active, since a small subset of families are responsible for most of the

polymorphisms. Interestingly, it seems like the LTR retrotransposons most active recently are different from those that were previously: the few families which are responsible for the currently polymorphic sites are all younger than 1MYa, less than the average of the LTR retro population, and notably after the peak activity 2MYa. .

The chromosomal distribution of fixed TEs, polymorphic insertions and genes attests to the equilibrium between the colonizing force of TEs and genomes' capacities for damage control. Recent TEs are frequently found near genes, while older ones are concentrated in heterochromatic regions, attesting to the process of selection against TE insertions near genes. This distribution has been observed in many species and suggests that centromeres are both a "haven" for TEs and that TEs might also be fulfilling a function there. Analysis of the *copia* and *gypsy* type retrotransposons shows that these different distributions are the result of two factors: insertion preference and selection. While there is a slight preference for gypsies to insert in heterochromatic regions, both types of elements seem to concentrate on heterochromatic regions with time, probably due to selection forces. To our knowledge this is the first analysis that uses polymorphic TEs to investigate differential distribution of recent and old transposons, and thus revealing at an intra-species scale the timeline of selection that leads to the final distributions observed.

The data provided in this work should give some insight into the TE activity during melon evolution and on when during this process TEs were more active. The analysis of the polymorphic sites with respect to the phylogenetic tree has allowed us to derive an insertion rate for each branch of the tree. We observe a higher insertion rate over the *agrestis* clade compared to *melo.* About a fourth of the PM sites between T111 and VED are in genes, and more than half of these in exons, and this proportion is holds true for the other varieties. Interestingly, is is the ancestral lineage to T111 and VED that holds the highest rate of TE expansion. This, combined with the fact that there is a high degree of DNA similarity between these two lines yet very different phenotypes, makes them ideal candidates to assess the importance of TEs in evolution.

As outlined in the introduction, TEs were first discovered due to the phenotype induced by insertional mutagenesis. In these cases, the impact of the TE was clear as it was inserted in a gene of known function, and the phenotype induced was what led to its discovery. Various elements in various genomes were thus identified, but with a certain ascertainment bias in regards to the overall impact of

TEs, as the only ones observed are those inducing a detectable phenotype. This led to the question of their prevalence and frequency of movement, and frequency with which this movement impacted gene expression or adaptation. Studies like the one we describe here show that now we have the tools to determine their prevalence in a genome and know how often polymorphisms occur, and identify hundreds of insertions in a genome where there is a potential TE effect directly on a gene. However, in order to interpret this – ask how often it has an impact – we need to know what these genes do. So though we can generate all these predictions computationally, we still need to have a good grasp on gene function to interpret this data. So here we come to a paradox: either we can observe the effect of a TE (by traditional genetics) or we can observe the genome-wide frequency of movements (by computational methods), but not both at the same time. In order to start making associations between TE-related polymorphisms and phenotype, in addition to more data as to individual genes' function, there is a need for highly systematic cataloging of phenotypes. In this respect projects such as the SolGenomics database (http://solgenomics.net/) which integrates both genomic and phenotypic data, organized in ontologies and for many Solanaceae species, become particularly useful. Assessing genome-wide the more subtle effects of TE polymorphisms such as epigenetic effects, small RNA-mediated silencing or transcription regulation becomes rapidly intractable since it is not limited to TE insertion in genes but can occur in *trans* and involves many other dimension of data such as expression profiles, epigenetic maps and small RNA profiles. These types of analyses are best attempted with a particular case in hand, and some clue that indicates a particular TE or family is a good candidate for study.

The study we performed in Arabidopsis is exactly one of these cases: we happened by chance (through the astuteness of a bioinformatician with the rare quality of actually manually perusing the output) on a TE with a potentially interesting profile: it contained the binding site for a master transcription factor, conserved across plants and animals and regulating key functions such as DNA replication and cell cycle. In this study we showed that the sequence for the E2F binding site is found in transposons in five related genomes: *Arabisopsis thaliana*, *Arabisopsis lyrata*, *Capsella rubella*, *Brassica rapa* and *Thelungela halophila*. It is found in various families, some shared by several (or all) genomes, and others genome-specific. These TEs, and the fact that the binding motif is included in larger, tandemly repeated motif, have amplified the number of occurrences of the E2F BS in all of these genomes except for *T. halophila*. Given the distance between the TE families that contain the E2F site,

and considering the mechanisms of creation, expansion and contraction of minisatellites, we can suppose various models to explain this phenomenon. It is still not clear what is the sequence of events that has led to the capture of this minisatellite by various TEs. Indeed, the minisatellite is much more conserved than the TEs that harbor it, rejecting the hypothesis that it was present in an ancestral TE. The sequence amplified corresponds to one of several that fits the consensus E2F binding site, and while it is bound by the protein, this cannot be the only selection mechanism (either for the transposon of for its function as a protein) otherwise we would have observed selection for other sequences that fit the consensus. Further investigation into the history of these elements and analysis of related genomes will help to elucidate this question

The impact of the presence of E2F sites in TEs is twofold: one, the impact of relocating these TFBS, and two, the impact of the amplification. There are a handful of insertions that potentially could affect gene expression: genes that are putative E2F targets and who's closest E2F BS is contributed by a TE. That the actualy E2F protein can bind to sites in transposons does not exclude the possibility that these sites could be fucntioning as regulatory sequences. This will not be the first time that a TE wires a new gene into a regulatory network, and it has actually been postulated that its the crucial TF like E2F or NANOG that need this combinatorial and randomization function, and that is what has permitted them to become master TF. Preliminary data indicates that these insertions are polymorphic between other varieties of *Arabidopsis thaliana*, indicating that any new genes wired into the E2F regulatory network by a TE would be recent, consistent with the trend that housekeeping genes do not contain TEs in their promoters, and those that do are recent and with specific functions (Van de Lagemaat et al. 2003)

Another potential effect is the titration of the protein away from its conventional targets. We first hypothesized that the E2F sites in TEs would not be accessible, hidden away in heterochromatin. Surprisingly, even though E2FBS in TEs do show heterochromatic epigenetic marks, they can be bound by the TF. However, this might not be the case in all tissues of in all stages of the cell cycle, and makes us wonder whether the genome has another way of regulating its binding.

Any advantage conferred to the MITEs by the E2F sequence must be limited to a particular family in the context of a particular genome, as the same families have amplified to different extents in differetn genomes.

This study contributes to the growing knowledge base of cases in which TEs have influenced gene expression, as vehicles for shuffling and translocating TFBS. As the complexity of gene regulation dictates the complexity of an organism, this particular function of TEs for generating combinations of TFBS and relocating them might be essential for evolution.

These two studies in melon and Arabidopsis taken together highlight the contradictory nature of transposable elements: on one hand, they are both invasive, expanding in bursts, and actively selected against and deleted from the genome, and on the other, in some occasions, are the source of essential innovations. How can something be potentially so advantageous, and as well so deleterious?

A model has been suggested by Hoen and Bureau (Douglas R Hoen, Thomas Bureau 2012) in which TEs can exist under two types of selection: replicative selection, or phenotypic selection. In general, they are under the former, and multiply in order to survive in the ecosystem which is their host genome. In general, this propensity is counteracted by selection against them, eliminating the more nefarious insertions. Occasionally, an individual TE insertion is advantageous, and now comes under phenotypic selection – this is the process of domestication. Like domesticated animals, the TE loses characteristics which are advantageous in the "wild" (such as mobility) and maintains only characteristics advantageous for its host (like a DNA binding motif). A balance is kept between the two, in which TEs are actually selected for their replicative nature – otherwise they would disappear – but also held in check by a slew of different mechanisms. They are kept around for those occasionally brilliant moments of innovative exaptation, when they fulfill a function that could never have been achieved through simple point mutations.

In the big picture, this dissertation work has contributed to the quest of finding how TEs impact plant genes and genomes. Studying both crop plants and model organisms leads to deepening our understanding of the plants that surround us and nourish us. Evolution of human societies is tightly entwined with that of the plants and animals we use for food and understanding them, the process of domestication and the mechanisms for diversification is essential in order to be able to adapt our food sources (and preserve the genetic diversity for them to adapt themselves) to a constantly changing world.

Answers tend to raise more questions, and particularly interesting directions to continue this work would be, on one hand, relating TE polymorphisms between melon lines to phenotypic traits, and on the other, determining by which mechanism the TEs in Arabidopsis and other Brassicas have captured the E2F binding site and to what extend this has modified E2F-regulation in those species. Finally, with the jitterbug tool in hand to detect TE polymorphisms and the wealth of genome sequences being generated, there is no dearth of TE variation data to be explored.

As the poet Rilke said:

"Try to love the questions themselves, like locked rooms and like books written in a foreign language." (Letters to a Young Poet)

Perhaps, genomes are like books in a foreign language and the immense number of organisms, species and ecosystems are like locked rooms … and our job as scientists is to love the questions.

References

Biessmann, H., K. Valgeirsdottir, A. Lofsky, C. Chin, B. Ginther, R. W. Levis, and M. L. Pardue. 1992. "HeT-A, a Transposable Element Specifically Involved in 'Healing' Broken Chromosome Ends in Drosophila Melanogaster." *Molecular and Cellular Biology* 12 (9) (September 1): 3910–3918. doi:10.1128/MCB.12.9.3910.

Bourque, Guillaume, Bernard Leong, Vinsensius B. Vega, Xi Chen, Yen Ling Lee, Kandhadayar G. Srinivasan, Joon-Lin Chew, et al. 2008. "Evolution of the Mammalian Transcription Factor Binding Repertoire via Transposable Elements." *Genome Research* (November 1). doi:10.1101/gr.080663.108. http://genome.cshlp.org/content/early/2008/10/03/gr.080663.108.

Butelli, Eugenio, Concetta Licciardello, Yang Zhang, Jianjun Liu, Steve Mackay, Paul Bailey, Giuseppe Reforgiato-Recupero, and Cathie Martin. 2012. "Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges[W][OA]." *The Plant Cell* 24 (3) (March): 1242–1255. doi:10.1105/tpc.111.095232.

Casola, Claudio, Donald Hucks, and Cedric Feschotte. 2008. "Convergent Domestication of Pogo-like Transposases into Centromere-binding Proteins in Fission Yeast and Mammals." *Molecular Biology and Evolution* 25 (1) (January): 29–41. doi:10.1093/molbev/msm221.

Cowley, Michael, and Rebecca J. Oakey. 2013. "Transposable Elements Re-Wire and Fine-Tune the Transcriptome." *PLoS Genet* 9 (1) (January 24): e1003234. doi:10.1371/journal.pgen.1003234.

Elbaidouri, Moaine, and Olivier Panaud. 2013. "Comparative Genomic Paleontology Across Plant Kingdom Reveals The Dynamics Of TE-driven Genome Evolution." *Genome Biology and Evolution* (February 20). doi:10.1093/gbe/evt025. http://gbe.oxfordjournals.org/content/early/2013/03/08/gbe.evt025.

Fedoroff, N., S. Wessler, and M. Shure. 1983. "Isolation of the Transposable Maize Controlling Elements Ac and Ds." *Cell* 35 (1): 235–242.

Feng, Gang, Young-Eun Leem, and Henry L. Levin. 2013. "Transposon Integration Enhances Expression of Stress Response Genes." *Nucleic Acids Research* 41 (2) (January): 775–789. doi:10.1093/nar/gks1185.

Fugmann, Sebastian D. 2010. "The Origins of the Rag genes—From Transposition to V (D) J Recombination." In *Seminars in Immunology*, 22:10–16. http://www.sciencedirect.com/science/article/pii/S1044532309001195.

Grandbastien, M A, A Spielmann, and M Caboche. 1989. "Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics." *Nature* 337 (6205) (January 26): 376–380. doi:10.1038/337376a0.

Hehl, R, W K Nacken, A Krause, H Saedler, and H Sommer. 1991. "Structural analysis of Tam3, a transposable element from Antirrhinum majus, reveals homologies to the Ac element from maize." *Plant molecular biology* 16 (2) (February): 369–371.

Hernández-Pinzón, Inmaculada, Marta Cifuentes, Elizabeth Hénaff, Néstor Santiago, M. Lluïsa Espinás, and Josep M. Casacuberta. 2012. "The Tnt1 Retrotransposon Escapes Silencing in Tobacco, Its Natural Host." *PLoS ONE* 7 (3) (March 30): e33816. doi:10.1371/journal.pone.0033816.

Hernández-Pinzón, Inmaculada, Erika de Jesús, Néstor Santiago, and Josep M Casacuberta. 2009. "The frequent transcriptional readthrough of the tobacco Tnt1 retrotransposon and its possible implications for the control of resistance genes." *Journal of molecular evolution* 68 (3) (March):

269–278. doi:10.1007/s00239-009-9204-y.
Hudson, Matthew E., Damon R. Lisch, and Peter H. Quail. 2003. “The FHY3 and FAR1 Genes Encode Transposase-related Proteins Involved in Regulation of Gene Expression by the Phytochrome A-signaling Pathway.” *The Plant Journal* 34 (4): 453–471. doi:10.1046/j.1365-313X.2003.01741.x.
Jin, Weiwei, Juliana R. Melo, Kiyotaka Nagaki, Paul B. Talbert, Steven Henikoff, R. Kelly Dawe, and Jiming Jiang. 2004. “Maize Centromeres: Organization and Functional Adaptation in the Genetic Background of Oat.” *The Plant Cell Online* 16 (3) (March 1): 571–581. doi:10.1105/tpc.018937.
Kapitonov, Vladimir V., and Jerzy Jurka. 2004. “Harbinger Transposons and an Ancient HARBI1 Gene Derived from a Transposase.” *DNA and Cell Biology* 23 (5): 311–324.
Khatun, Jainab, Sarah Djebali, Carrie A. Davis, Angelika Merkel, and Thomas R. Gingeras. 2012. “Landscape of Transcription in Human Cells.” *Nature*. http://scholarworks.boisestate.edu/bio_facpubs/260/.
Kunarso, Galih, Na-Yu Chia, Justin Jeyakani, Catalina Hwang, Xinyi Lu, Yun-Shen Chan, Huck-Hui Ng, and Guillaume Bourque. 2010. “Transposable Elements Have Rewired the Core Regulatory Network of Human Embryonic Stem Cells.” *Nature Genetics* 42 (7): 631–634.
McClintock, Barbara. 1983. “The Significance of Responses of the Genome to Challenge.” *Physiology Or Medicine Literature Peace Economic Sciences*: 180.
McGinnis, W, A W Shermoen, and S K Beckendorf. 1983. “A transposable element inserted just 5’ to a Drosophila glue protein gene alters gene expression and chromatin structure.” *Cell* 34 (1) (August): 75–84.
Mikkelsen, Tarjei S., Matthew J. Wakefield, Bronwen Aken, Chris T. Amemiya, Jean L. Chang, Shannon Duke, Manuel Garber, et al. 2007. “Genome of the Marsupial Monodelphis Domestica Reveals Innovation in Non-coding Sequences.” *Nature* 447 (7141) (May 10): 167–177. doi:10.1038/nature05805.
Momose, Masaki, Yutaka Abe, and Yoshihiro Ozeki. 2010. “Miniature Inverted-Repeat Transposable Elements of Stowaway Are Active in Potato.” *Genetics* 186 (1) (September): 59 –66. doi:10.1534/genetics.110.117606.
Naito, Ken, Eunyoung Cho, Guojun Yang, Matthew A. Campbell, Kentaro Yano, Yutaka Okumoto, Takatoshi Tanisaka, and Susan R. Wessler. 2006. “Dramatic Amplification of a Rice Transposable Element During Recent Domestication.” *Proceedings of the National Academy of Sciences* 103 (47) (November 21): 17620–17625. doi:10.1073/pnas.0605421103.
Nakazaki, Tetsuya, Yutaka Okumoto, Akira Horibata, Satoshi Yamahira, Masayoshi Teraishi, Hidetaka Nishida, Hiromo Inoue, and Takatoshi Tanisaka. 2003. “Mobilization of a Transposon in the Rice Genome.” *Nature* 421 (6919) (January 9): 170–172. doi:10.1038/nature01219.
Nosaka, Misuzu, Jun-Ichi Itoh, Yasuo Nagato, Akemi Ono, Aiko Ishiwata, and Yutaka Sato. 2012. “Role of Transposon-Derived Small RNAs in the Interplay Between Genomes and Parasitic DNA in Rice.” *PLoS Genet* 8 (9) (September 27): e1002953. doi:10.1371/journal.pgen.1002953.
O’Hare, K, and G M Rubin. 1983. “Structures of P transposable elements and their sites of insertion and excision in the Drosophila melanogaster genome.” *Cell* 34 (1) (August): 25–35.
Pecinka, Ales, Huy Q. Dinh, Tuncay Baubec, Marisa Rosa, Nicole Lettner, and Ortrun Mittelsten Scheid. 2010. “Epigenetic Regulation of Repetitive Elements Is Attenuated by Prolonged Heat Stress in Arabidopsis.” *The Plant Cell Online* 22 (9): 3118 –3129. doi:10.1105/tpc.110.078493.
Pereira, Andy, Zsuzsanna Schwarz-Sommer, Alfons Gierl, Isolde Bertram, Peter A. Peterson, and Heinz

Saedler. 1985. “Genetic and Molecular Analysis of the Enhancer (En) Transposable Element System of Zea Mays.” *The EMBO Journal* 4 (1) (January): 17–23.
Piriyapongsa, Jittima, and I. King Jordan. 2007. “A Family of Human MicroRNA Genes from Miniature Inverted-Repeat Transposable Elements.” *PLoS ONE* 2 (2) (February 14): e203. doi:10.1371/journal.pone.0000203.
Piriyapongsa, Jittima, and I. King Jordan. 2008. “Dual Coding of siRNAs and miRNAs by Plant Transposable Elements.” *RNA* 14 (5) (May): 814–821. doi:10.1261/rna.916708.
Plasterk, Ronald H. A. 2002. “RNA Silencing: The Genome’s Immune System.” *Science* 296 (5571) (May 17): 1263–1265. doi:10.1126/science.1072148.
Salem, Abdel-Halim, David A. Ray, Jinchuan Xing, Pauline A. Callinan, Jeremy S. Myers, Dale J. Hedges, Randall K. Garber, David J. Witherspoon, Lynn B. Jorde, and Mark A. Batzer. 2003. “Alu Elements and Hominid Phylogenetics.” *Proceedings of the National Academy of Sciences* 100 (22) (October 28): 12787–12791. doi:10.1073/pnas.2133766100.
Thurman, Robert E., Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield, et al. 2012. “The Accessible Chromatin Landscape of the Human Genome.” *Nature* 489 (7414) (September 6): 75–82. doi:10.1038/nature11232.
Venter, J. Craig, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, et al. 2001. “The Sequence of the Human Genome.” *Science* 291 (5507) (February 16): 1304 –1351. doi:10.1126/science.1058040.
Waterhouse, Peter M., Ming-Bo Wang, and Tony Lough. 2001. “Gene Silencing as an Adaptive Defence Against Viruses.” *Nature* 411 (6839) (June 14): 834–842. doi:10.1038/35081168.
Wicker, Thomas, Francois Sabot, Aurelie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2007. “A Unified Classification System for Eukaryotic Transposable Elements.” *Nat Rev Genet* 8 (12) (December): 973–982. doi:10.1038/nrg2165.
Wicker, Thomas, Stefan Taudien, Andreas Houben, Beat Keller, Andreas Graner, Matthias Platzer, and Nils Stein. 2009. “A Whole-genome Snapshot of 454 Sequences Exposes the Composition of the Barley Genome and Provides Evidence for Parallel Evolution of Genome Size in Wheat and Barley.” *The Plant Journal* 59 (5): 712–722. doi:10.1111/j.1365-313X.2009.03911.x.

Results Chapter 1

Argout, Xavier, Jerome Salse, Jean-Marc Aury, Mark J. Guiltinan, Gaetan Droc, Jerome Gouzy, Mathilde Allegre, et al. 2011. “The Genome of Theobroma Cacao.” *Nature Genetics* 43 (2) (February): 101–108. doi:10.1038/ng.736.
Bergman, Casey M., and Hadi Quesneville. 2007. “Discovering and Detecting Transposable Elements in Genome Sequences.” *Briefings in Bioinformatics* 8 (6) (November 1): 382–392. doi:10.1093/bib/bbm048.
Buisine, Nicolas, Hadi Quesneville, and Vincent Colot. 2008. “Improved Detection and Annotation of Transposable Elements in Sequenced Genomes Using Multiple Reference Sequence Sets.” *Genomics* 91 (5) (May): 467–475. doi:10.1016/j.ygeno.2008.01.005.
Chen, Yong, Fengfeng Zhou, Guojun Li, and Ying Xu. 2009. “MUST: A System for Identification of Miniature Inverted-repeat Transposable Elements and Applications to Anabaena Variabilis and Haloquadratum Walsbyi.” *Gene* 436 (1-2) (May 1): 1–7. doi:10.1016/j.gene.2009.01.019.
Choulet, Frédéric, Thomas Wicker, Camille Rustenholz, Etienne Paux, Jérome Salse, Philippe Leroy,

Stéphane Schlub, et al. 2010. “Megabase Level Sequencing Reveals Contrasted Organization and Evolution Patterns of the Wheat Gene and Transposable Element Spaces.” *The Plant Cell Online* 22 (6) (June 1): 1686–1701. doi:10.1105/tpc.110.074187.

Devos, Katrien M., James K. M. Brown, and Jeffrey L. Bennetzen. 2002. “Genome Size Reduction Through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis.” *Genome Research* 12 (7) (July 1): 1075–1079. doi:10.1101/gr.132102.

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5) (March 1): 1792–1797. doi:10.1093/nar/gkh340.

———. 2010. “Search and Clustering Orders of Magnitude Faster Than BLAST.” *Bioinformatics*. doi:10.1093/bioinformatics/btq461. http://bioinformatics.oxfordjournals.org/content/early/2010/08/12/bioinformatics.btq461.abstract.

Flutre, Timothée, Elodie Duprat, Catherine Feuillet, and Hadi Quesneville. 2011. “Considering Transposable Element Diversification in De Novo Annotation Approaches.” *PLoS ONE* 6 (1) (January 31): e16526. doi:10.1371/journal.pone.0016526.

Garcia-Mas, Jordi, Andrej Benjak, Walter Sanseverino, Michael Bourgeois, Gisela Mir, Víctor M. González, Elizabeth Hénaff, et al. 2012. “The Genome of Melon (Cucumis Melo L.).” *Proceedings of the National Academy of Sciences* (July 2). doi:10.1073/pnas.1205415109. http://www.pnas.org/content/early/2012/06/28/1205415109.

Gonzalez, Victor, Andrej Benjak, Elizabeth Henaff, Gisela Mir, Josep Casacuberta, Jordi Garcia-Mas, and Pere Puigdomenech. 2010. “Sequencing of 6.7 Mb of the Melon Genome Using a BAC Pooling Strategy.” *BMC Plant Biology* 10 (1): 246. doi:10.1186/1471-2229-10-246.

Guermonprez, Hélène, Elizabeth Hénaff, Marta Cifuentes, and Josep Casacuberta. 2013. “MITEs, Miniature Elements with a Major Role in Plant Genome Evolution - Springer.” Accessed April 15. http://link.springer.com/chapter/10.1007%2F978-3-642-31842-9_7#page-1.

Han, Yujun, and Susan R. Wessler. 2010. “MITE-Hunter: a Program for Discovering Miniature Inverted-repeat Transposable Elements from Genomic Sequences.” *Nucleic Acids Research*. doi:10.1093/nar/gkq862. http://nar.oxfordjournals.org/content/early/2010/09/29/nar.gkq862.abstract.

Hollister, Jesse D., and Brandon S. Gaut. 2009. “Epigenetic Silencing of Transposable Elements: A Trade-off Between Reduced Transposition and Deleterious Effects on Neighboring Gene Expression.” *Genome Research* 19 (8) (August): 1419–1428. doi:10.1101/gr.091678.109.

Huang, Sanwen, Ruiqiang Li, Zhonghua Zhang, Li Li, Xingfang Gu, Wei Fan, William J Lucas, et al. 2009. “The Genome of the Cucumber, Cucumis Sativus L.” *Nat Genet* 41 (12) (December): 1275–1281. doi:10.1038/ng.475.

Juretic, Nikoleta, Thomas E. Bureau, and Richard M. Bruskiewich. 2004. “Transposable Element Annotation of the Rice Genome.” *Bioinformatics* 20 (2) (January 22): 155–160. doi:10.1093/bioinformatics/bth019.

Li, Xuehui, Tamer Kahveci, and A. Mark Settles. 2008. “A Novel Genome-scale Repeat Finder Geared Towards Transposons.” *Bioinformatics* 24 (4) (February 15): 468 –476. doi:10.1093/bioinformatics/btm613.

Lockton, Steven, Jeffrey Ross-Ibarra, and Brandon S. Gaut. 2008. “Demography and Weak Selection Drive Patterns of Transposable Element Diversity in Natural Populations of Arabidopsis Lyrata.” *Proceedings of the National Academy of Sciences of the United States of America* 105

(37) (September 16): 13965–13970. doi:10.1073/pnas.0804671105.
Naito, Ken, Eunyoung Cho, Guojun Yang, Matthew A. Campbell, Kentaro Yano, Yutaka Okumoto, Takatoshi Tanisaka, and Susan R. Wessler. 2006. “Dramatic Amplification of a Rice Transposable Element During Recent Domestication.” *Proceedings of the National Academy of Sciences* 103 (47) (November 21): 17620–17625. doi:10.1073/pnas.0605421103.
Price, Alkes L., Neil C. Jones, and Pavel A. Pevzner. 2005. “De Novo Identification of Repeat Families in Large Genomes.” *Bioinformatics* 21 (suppl 1) (June 1): i351–i358. doi:10.1093/bioinformatics/bti1018.
Quesneville, Hadi, Casey M Bergman, Olivier Andrieu, Delphine Autard, Danielle Nouaud, Michael Ashburner, and Dominique Anxolabehere. 2005. “Combined Evidence Annotation of Transposable Elements in Genome Sequences.” *PLoS Computational Biology* 1 (2) (July). doi:10.1371/journal.pcbi.0010022.
Schnable, Patrick S, Doreen Ware, Robert S Fulton, Joshua C Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, et al. 2009. “The B73 maize genome: complexity, diversity, and dynamics.” *Science (New York, N.Y.)* 326 (5956) (November 20): 1112–1115. doi:10.1126/science.1178534.
Venter, J. Craig, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, et al. 2001. “The Sequence of the Human Genome.” *Science* 291 (5507) (February 16): 1304 –1351. doi:10.1126/science.1058040.
Vitte, Clémentine, Olivier Panaud, and Hadi Quesneville. 2007. “LTR Retrotransposons in Rice (Oryza Sativa, L.): Recent Burst Amplifications Followed by Rapid DNA Loss.” *BMC Genomics* 8 (1) (July 6): 218. doi:10.1186/1471-2164-8-218.
Vitte, C, and O Panaud. 2005. “LTR Retrotransposons and Flowering Plant Genome Size: Emergence of the Increase/decrease Model.” *Cytogenetic and Genome Research* 110 (1-4): 91–107. doi:10.1159/000084941.
Wang, Hao, and Jin-Song Liu. 2008. “LTR Retrotransposon Landscape in Medicago Truncatula: More Rapid Removal Than in Rice.” *BMC Genomics* 9 (August 10): 382. doi:10.1186/1471-2164-9-382.
Xu, Zhao, and Hao Wang. 2007. “LTR_FINDER: An Efficient Tool for the Prediction of Full-length LTR Retrotransposons.” *Nucleic Acids Research* 35 (suppl 2) (July 1): W265–W268. doi:10.1093/nar/gkm286.

Results Chapter 2

Ahmed, Ikhlak, Alexis Sarazin, Chris Bowler, Vincent Colot, and Hadi Quesneville. 2011. “Genome-wide Evidence for Local DNA Methylation Spreading from Small RNA-targeted Sequences in Arabidopsis.” *Nucleic Acids Research* (May 17). doi:10.1093/nar/gkr324. http://nar.oxfordjournals.org/content/early/2011/05/17/nar.gkr324.
Alkan, Can, Bradley P. Coe, and Evan E. Eichler. 2011. “Genome Structural Variation Discovery and Genotyping.” *Nature Reviews Genetics* 12 (5) (May 1): 363–376. doi:10.1038/nrg2958.
Beck, Christine R., Pamela Collier, Catriona Macfarlane, Maika Malig, Jeffrey M. Kidd, Evan E. Eichler, Richard M. Badge, and John V. Moran. 2010. “LINE-1 Retrotransposition Activity in

Human Genomes." *Cell* 141 (7) (June 25): 1159–1170. doi:10.1016/j.cell.2010.05.021.
Ewing, Adam D., and Haig H. Kazazian. 2010. "High-throughput Sequencing Reveals Extensive Variation in Human-specific L1 Content in Individual Human Genomes." *Genome Research* 20 (9) (September): 1262–1270. doi:10.1101/gr.106419.110.
Freeling, Michael, Margaret R Woodhouse, Shabarinath Subramaniam, Gina Turco, Damon Lisch, and James C Schnable. 2012. "Fractionation Mutagenesis and Similar Consequences of Mechanisms Removing Dispensable or Less-expressed DNA in Plants." *Current Opinion in Plant Biology* 15 (2) (April): 131–139. doi:10.1016/j.pbi.2012.01.015.
Hajirasouliha, Iman, Fereydoun Hormozdiari, Can Alkan, Jeffrey M. Kidd, Inanc Birol, Evan E. Eichler, and S. Cenk Sahinalp. 2010. "Detection and Characterization of Novel Sequence Insertions Using Paired-end Next-generation Sequencing." *Bioinformatics* 26 (10) (May 15): 1277 –1283. doi:10.1093/bioinformatics/btq152.
Hormozdiari, Fereydoun, Iman Hajirasouliha, Phuong Dao, Faraz Hach, Deniz Yorukoglu, Can Alkan, Evan E. Eichler, and S. Cenk Sahinalp. 2010. "Next-generation VariationHunter: Combinatorial Algorithms for Transposon Insertion Discovery." *Bioinformatics* 26 (12) (June 15): i350–i357. doi:10.1093/bioinformatics/btq216.
Huang, Xuehui, Guojun Lu, Qiang Zhao, Xiaohui Liu, and Bin Han. 2008. "Genome-Wide Analysis of Transposon Insertion Polymorphisms Reveals Intraspecific Variation in Cultivated Rice." *Plant Physiology* 148 (1) (September 1): 25–40. doi:10.1104/pp.108.121491.
Joly-Lopez, Zoé, Ewa Forczek, Douglas R. Hoen, Nikoleta Juretic, and Thomas E. Bureau. 2012. "A Gene Family Derived from Transposable Elements During Early Angiosperm Evolution Has Reproductive Fitness Benefits in Arabidopsis Thaliana." *PLoS Genet* 8 (9) (September 6): e1002931. doi:10.1371/journal.pgen.1002931.
Lee, Eunjung, Rebecca Iskow, Lixing Yang, Omer Gokcumen, Psalm Haseley, Lovelace J. Luquette, Jens G. Lohr, et al. 2012. "Landscape of Somatic Retrotransposition in Human Cancers." *Science* (June 28). doi:10.1126/science.1222077. http://www.sciencemag.org/content/early/2012/06/27/science.1222077.
Lee, Hayan, and Michael C. Schatz. 2012. "Genomic Dark Matter: The Reliability of Short Read Mapping Illustrated by the Genome Mappability Score." *Bioinformatics* (June 4). doi:10.1093/bioinformatics/bts330. http://bioinformatics.oxfordjournals.org/content/early/2012/06/04/bioinformatics.bts330.
Lisch, Damon. 2013. "How Important Are Transposons for Plant Evolution?" *Nature Reviews Genetics* 14 (1) (January): 49–61. doi:10.1038/nrg3374.
Ray, David A., and Mark A. Batzer. 2011. "Reading TE Leaves: New Approaches to the Identification of Transposable Element Insertions." *Genome Research* 21 (6) (June): 813–820. doi:10.1101/gr.110528.110.
Sabot, François, Nathalie Picault, Moaine El-Baidouri, Christel Llauro, Cristian Chaparro, Benoit Piegu, Anne Roulin, et al. 2011. "Transpositional Landscape of the Rice Genome Revealed by Paired-end Mapping of High-throughput Re-sequencing Data." *The Plant Journal* 66 (2) (April 1): 241–246. doi:10.1111/j.1365-313X.2011.04492.x.
Stewart, Chip, Deniz Kural, Michael P. Strömberg, Jerilyn A. Walker, Miriam K. Konkel, Adrian M. Stütz, Alexander E. Urban, et al. 2011. "A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans." *PLoS Genet* 7 (8): e1002236. doi:10.1371/journal.pgen.1002236.
Wang, Hao, and Jin-Song Liu. 2008. "LTR Retrotransposon Landscape in Medicago Truncatula: More

Rapid Removal Than in Rice." *BMC Genomics* 9 (August 10): 382. doi:10.1186/1471-2164-9-382.

Wang, Xi, Detlef Weigel, and Lisa M. Smith. 2013. "Transposon Variants and Their Effects on Gene Expression in Arabidopsis." *PLoS Genet* 9 (2) (February 7): e1003255. doi:10.1371/journal.pgen.1003255.

Ye, Kai, Marcel H. Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. 2009. "Pindel: a Pattern Growth Approach to Detect Break Points of Large Deletions and Medium Sized Insertions from Paired-end Short Reads." *Bioinformatics* 25 (21) (November 1): 2865–2871. doi:10.1093/bioinformatics/btp394.

Results Chapter 3

Benson, Gary. 1999. "Tandem Repeats Finder: a Program to Analyze DNA Sequences." *Nucleic Acids Research* 27 (2) (January 1): 573–580. doi:10.1093/nar/27.2.573.

Bourque, Guillaume, Bernard Leong, Vinsensius B. Vega, Xi Chen, Yen Ling Lee, Kandhadayar G. Srinivasan, Joon-Lin Chew, et al. 2008. "Evolution of the Mammalian Transcription Factor Binding Repertoire via Transposable Elements." *Genome Research* (November 1). doi:10.1101/gr.080663.108. http://genome.cshlp.org/content/early/2008/10/03/gr.080663.108.

Cui, Feng, Michael V Sirotin, and Victor B Zhurkin. 2011. "Impact of Alu Repeats on the Evolution of Human P53 Binding Sites." *Biology Direct* 6 (January 6): 2. doi:10.1186/1745-6150-6-2.

Gifford, Wesley D., Samuel L. Pfaff, and Todd S. Macfarlan. 2013. "Transposable Elements as Genetic Regulatory Substrates in Early Development." *Trends in Cell Biology*. doi:10.1016/j.tcb.2013.01.001. http://www.sciencedirect.com/science/article/pii/S0962892413000032.

Hudson, Matthew E., Damon R. Lisch, and Peter H. Quail. 2003. "The FHY3 and FAR1 Genes Encode Transposase-related Proteins Involved in Regulation of Gene Expression by the Phytochrome A-signaling Pathway." *The Plant Journal* 34 (4): 453–471. doi:10.1046/j.1365-313X.2003.01741.x.

Hu, Tina T., Pedro Pattyn, Erica G. Bakker, Jun Cao, Jan-Fang Cheng, Richard M. Clark, Noah Fahlgren, et al. 2011. "The Arabidopsis Lyrata Genome Sequence and the Basis of Rapid Genome Size Change." *Nature Genetics* 43 (5) (May): 476–481. doi:10.1038/ng.807.

Kunarso, Galih, Na-Yu Chia, Justin Jeyakani, Catalina Hwang, Xinyi Lu, Yun-Shen Chan, Huck-Hui Ng, and Guillaume Bourque. 2010. "Transposable Elements Have Rewired the Core Regulatory Network of Human Embryonic Stem Cells." *Nat Genet* 42 (7) (July): 631–634. doi:10.1038/ng.600.

Liu, Jun, Yuehui He, Richard Amasino, and Xuemei Chen. 2004. "siRNAs Targeting an Intronic Transposon in the Regulation of Natural Flowering Behavior in Arabidopsis." *Genes & Development* 18 (23) (December 1): 2873–2878. doi:10.1101/gad.1217304.

Lowe, Craig B., and David Haussler. 2012. "29 Mammalian Genomes Reveal Novel Exaptations of Mobile Elements for Likely Regulatory Functions in the Human Genome." *PLoS ONE* 7 (8) (August 27): e43128. doi:10.1371/journal.pone.0043128.

Micale, Lucia, Maria Nicla Loviglio, Marta Manzoni, Carmela Fusco, Bartolomeo Augello, Eugenia Migliavacca, Grazia Cotugno, et al. 2012. "A Fish-Specific Transposable Element Shapes the Repertoire of P53 Target Genes in Zebrafish." *PLoS ONE* 7 (10) (October 31): e46642. doi:10.1371/journal.pone.0046642.

Miele, Vincent, Simon Penel, and Laurent Duret. 2011. “Ultra-fast Sequence Clustering from Similarity Networks with SiLiX.” *BMC Bioinformatics* 12 (1) (April 22): 116. doi:10.1186/1471-2105-12-116.

Mikkelsen, Tarjei S., Matthew J. Wakefield, Bronwen Aken, Chris T. Amemiya, Jean L. Chang, Shannon Duke, Manuel Garber, et al. 2007. “Genome of the Marsupial Monodelphis Domestica Reveals Innovation in Non-coding Sequences.” *Nature* 447 (7141) (May 10): 167–177. doi:10.1038/nature05805.

Naouar, Naïra, Klaas Vandepoele, Tim Lammens, Tineke Casneuf, Georg Zeller, Paul Van Hummelen, Detlef Weigel, et al. 2008. “Quantitative RNA Expression Analysis with Affymetrix Tiling 1.0R Arrays Identifies New E2F Target Genes.” *The Plant Journal* 57 (1) (September 19): 184–194. doi:10.1111/j.1365-313X.2008.03662.x.

Ramirez-Parra, Elena, Corinne Fründt, and Crisanto Gutierrez. 2003. “A Genome-wide Identification of E2F-regulated Genes in Arabidopsis.” *The Plant Journal* 33 (4): 801–811. doi:10.1046/j.1365-313X.2003.01662.x.

Ramirez-Parra, Elena, M. Angeles López-Matas, Corinne Fründt, and Crisanto Gutierrez. 2004. “Role of an Atypical E2F Transcription Factor in the Control of Arabidopsis Cell Growth and Differentiation.” *The Plant Cell Online* 16 (9): 2350 –2363. doi:10.1105/tpc.104.023978.

Richard, Guy-Franck, Alix Kerrest, and Bernard Dujon. 2008. “Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes.” *Microbiology and Molecular Biology Reviews* 72 (4) (December): 686 –727. doi:10.1128/MMBR.00011-08.

Testori, Alessandro, Livia Caizzi, Santina Cutrupi, Olivier Friard, Michele De Bortoli, Davide Cora’, and Michele Caselle. 2012. “The Role of Transposable Elements in Shaping the Combinatorial Interaction of Transcription Factors.” *BMC Genomics* 13 (1) (August 16): 400. doi:10.1186/1471-2164-13-400.

Vandepoele, Klaas, Kobe Vlieghe, Kobe Florquin, Lars Hennig, Gerrit T.S. Beemster, Wilhelm Gruissem, Yves Van de Peer, Dirk Inzé, and Lieven De Veylder. 2005. “Genome-Wide Identification of Potential Plant E2F Target Genes.” *Plant Physiology* 139 (1) (September): 316–328. doi:10.1104/pp.105.066290.

De Veylder, Lieven, Tom Beeckman, Gerrit T S Beemster, Janice de Almeida Engler, Sandra Ormenese, Sara Maes, Mirande Naudts, et al. 2002. “Control of Proliferation, Endoreduplication and Differentiation by the Arabidopsis E2Fa-DPa Transcription Factor.” *The EMBO Journal* 21 (6) (March 15): 1360–1368. doi:10.1093/emboj/21.6.1360.

Vinces, Marcelo D., Matthieu Legendre, Marina Caldara, Masaki Hagihara, and Kevin J. Verstrepen. 2009. “Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability.” *Science* 324 (5931) (May 29): 1213 –1216. doi:10.1126/science.1170097.

Wang, Ting, Jue Zeng, Craig B. Lowe, Robert G. Sellers, Sofie R. Salama, Min Yang, Shawn M. Burgess, Rainer K. Brachmann, and David Haussler. 2007. “Species-specific Endogenous Retroviruses Shape the Transcriptional Network of the Human Tumor Suppressor Protein P53.” *Proceedings of the National Academy of Sciences* 104 (47) (November 20): 18613–18618. doi:10.1073/pnas.0703637104.

Discussion

Douglas R Hoen, Thomas E. Bureau. 2012. “Transposable Element Exaptation in Plants”: 219–251.

doi:10.1007/978-3-642-31842-9_12.
Van de Lagemaat, Louie N, Josette-Renée Landry, Dixie L Mager, and Patrik Medstrand. 2003. "Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions." *Trends in genetics: TIG* 19 (10) (October): 530–536. doi:10.1016/j.tig.2003.08.004.

Annex

BMC Plant Biology

RESEARCH ARTICLE Open Access

Sequencing of 6.7 Mb of the melon genome using a BAC pooling strategy

Víctor M González[1†], Andrej Benjak[2†], Elizabeth Marie Hénaff[1], Gisela Mir[2], Josep M Casacuberta[1], Jordi Garcia-Mas[2], Pere Puigdomènech[1*]

Abstract

Background: *Cucumis melo* (melon) belongs to the Cucurbitaceae family, whose economic importance among horticulture crops is second only to Solanaceae. Melon has a high intra-specific genetic variation, morphologic diversity and a small genome size (454 Mb), which make it suitable for a great variety of molecular and genetic studies. A number of genetic and genomic resources have already been developed, such as several genetic maps, BAC genomic libraries, a BAC-based physical map and EST collections. Sequence information would be invaluable to complete the picture of the melon genomic landscape, furthering our understanding of this species' evolution from its relatives and providing an important genetic tool. However, to this day there is little sequence data available, only a few melon genes and genomic regions are deposited in public databases. The development of massively parallel sequencing methods allows envisaging new strategies to obtain long fragments of genomic sequence at higher speed and lower cost than previous Sanger-based methods.

Results: In order to gain insight into the structure of a significant portion of the melon genome we set out to perform massive sequencing of pools of BAC clones. For this, a set of 57 BAC clones from a double haploid line was sequenced in two pools with the 454 system using both shotgun and paired-end approaches. The final assembly consists of an estimated 95% of the actual size of the melon BAC clones, with most likely complete sequences for 50 of the BACs, and a total sequence coverage of 39x. The accuracy of the assembly was assessed by comparing the previously available Sanger sequence of one of the BACs against its 454 sequence, and the polymorphisms found involved only 1.7 differences every 10,000 bp that were localized in 15 homopolymeric regions and two dinucleotide tandem repeats. Overall, the study provides approximately 6.7 Mb or 1.5% of the melon genome. The analysis of this new data has allowed us to gain further insight into characteristics of the melon genome such as gene density, average protein length, or microsatellite and transposon content. The annotation of the BAC sequences revealed a high degree of collinearity and protein sequence identity between melon and its close relative *Cucumis sativus* (cucumber). Transposon content analysis of the syntenic regions suggests that transposition activity after the split of both cucurbit species has been low in cucumber but very high in melon.

Conclusions: The results presented here show that the strategy followed, which combines shotgun and BAC-end sequencing together with anchored marker information, is an excellent method for sequencing specific genomic regions, especially from relatively compact genomes such as that of melon. However, in agreement with other results, this map-based, BAC approach is confirmed to be an expensive way of sequencing a whole plant genome. Our results also provide a partial description of the melon genome's structure. Namely, our analysis shows that the melon genome is highly collinear with the smaller one of cucumber, the size difference being mainly due to the expansion of intergenic regions and proliferation of transposable elements.

* Correspondence: pprgmp@cid.csic.es
† Contributed equally
[1]Molecular Genetics Department, Center for Research in Agricultural Genomics CRAG (CSIC-IRTA-UAB), Jordi Girona, 18-26, 08034 Barcelona, Spain
Full list of author information is available at the end of the article

Background

During recent years an important effort has been made to increase the tools available for the genomic analysis of major plant crop species. Since the first genome sequence available of *Arabidopsis thaliana* [1], several others have been published. They include model plants such as *Brachypodium* [2] but, increasingly, species that have been chosen for their importance in agriculture. For example the rice [3], maize [4], sorghum [5] or soybean [6] genomes are complex but the wealth of genetic information matches their economic interest. Consequently, for both scientific and economic reasons an increasing number of plant genomes are being analyzed, providing important resources useful for their biological study and breeding.

Several species of interest from both scientific and economic perspectives are of the Cucurbitaceae family. These include melon, cucumber, watermelon and squashes, all of which have been the object of biological and agricultural interest for centuries. In recent years various molecular tools have been established. For instance, the first assembly of the cucumber genome [7], as well as an increasing number of genetic and genomic resources developed for melon, a diploid species with a relatively compact (around 454 Mb [8]) genome [9]. These include tools such as a collection of more than 129,000 ESTs [10,11], BAC libraries [12,13], oligo-based microarrays [14,15], TILLING and EcoTILLING platforms [16,17], a set of near isogenic lines (NILs) [18] and several melon genetic maps [11,19-25]. Recently, we have built a physical map with 0.9x genomic coverage using both a BAC library and a genetic map previously developed in our laboratories [http://melonomics.upv.es/public_files, [26]], the first report of such a genomic resource of a Cucurbitaceae species so far. This physical map has also been integrated with the genetic map by anchoring a number of physical contigs (representing 12% of the melon genome) to 175 known genetic markers. These tools have been useful in the study of interesting agronomical traits such as virus or fungi resistance [27,28], sex determination [29,30] or the control of ripening [31,32]. These results demonstrate that molecular genetic approaches can successfully be used in melon to address basic questions of biological or agronomic relevance.

More extensive sequence information would be invaluable to complete the picture of the melon genomic landscape. Indeed, the sequences of only a few selected genomic regions have been published, totaling no more than 500 kb [29,33-35] and as of May 2010 no more than 173 melon genes can be found in GenBank [11], although a collection of ESTs probably representing more than 70% of the transcriptome is currently available [11]. The sequencing of the Sorghum genome has shown the feasibility of sequencing a plant genome larger than that of melon (730 Mb) using a Sanger-based whole genome shotgun approach [5]. However, the development of new massively parallel sequencing technologies allows envisaging a complete sequencing of the species at higher speed and at lower cost than previous Sanger-based methods. To this end, both whole genome sequencing approaches as well as map-based, BAC-to-BAC strategies have been proposed to sequence plant genomes [36,37].

A small number of research projects involving 454 sequencing of BAC clones have currently been published. In a pioneering study aimed at analyzing how 454 technology would perform on template derived from large genomes rich in repetitive content, four barley BAC clones 102-120 kb long, two of which had been previously sequenced using Sanger technology, were sequenced using 454 [38]. The results showed that gene-containing regions could efficiently and accurately be assembled into contigs, even at read coverages as low as x10.

In a later work eight BACs belonging to a minimum tiling path covering *ca*. 1 Mb of the Atlantic salmon genome were sequenced using 454 technology, the first published use of paired-end reads for *de novo* sequence assembly [39]. This study demonstrated that although the inclusion of paired-end reads greatly improved sequence assembly, there remained a significant number of gaps when compared to Sanger-generated sequencing data. Thus the authors concluded that, when it comes to *de novo* sequencing complex genomes, 454 sequencing should be restricted, at least for the time being, to establishing a set of ordered sequenced contigs.

Although these studies show that 454 sequencing can be used to assemble gene-containing regions from genomic sequences using a BAC-to-BAC approach, the cost of 454 sequencing individual BACs has led to consider pooling individual clones as a means to increase throughput and reduce the cost of genome sequencing. In one published study, 166 BACs totalling 20 Mb were divided into six pools of overlapping BACs, aided by paired-end sequencing. These were then used to 454-sequence a minimum tiling path which covered an entire chromosome arm from *Oryza barthii* [37]. The report shows that pooling BACs does not increase the complexity to a degree that makes assembly impossible, what makes this approach a feasible strategy for reducing the cost of BAC sequencing. In another work 91 barley BAC clones, pooled by sets of 12 or 24, were sequenced using 454 technology [40]. The introduction of short sequence tags to fragmented BAC DNA prior to pooling and sequencing helped to resolve the assembly of multiplex sequencing data by establishing

relationships between BAC clones and sequence reads, reducing sample complexity.

Here we present a pilot project aiming to sequence two pools of 35 and 23 melon BACs using the 454 system and a combination of shotgun and paired-end sequencing. The goal of the study was twofold: obtain sequence data for a significant proportion of the melon genome and thus insight into its structure, and test the strategy of massively sequencing pools of BACs. The results obtained allow an accuracy assessment of 454 sequence and assembly data as compared with sequence data produced using classical Sanger technology. Overall, the study provides approximately 7 Mb or 1.5% of the melon genome as a first step towards the complete sequence. The analysis of this data has provided insight into characteristics of the melon genome such as gene density, transposon content and synteny with cucumber.

Results and discussion

Selection of BAC clones for pooling and sequencing

Two pools of DNA prepared from BACs were sequenced using the 454 pyrosequencing method. These BACs had been produced from DNA of the double haploid line PIT92 obtained from the cross of PI 161375 and T111 as described in [12].

A set of 178 genetic markers selected from previous versions of the PI 161375 × T111 melon genetic map (mainly RFLPs [21] and SNPs [24,31,41,42]) were used to anchor 845 BAC clones from our genomic library to the genetic map [26]. Of these, a batch of 32 BACs anchored to genetic markers distributed throughout the genome (See Figure 1) were chosen for 454-sequencing. In order to test the quality of the sequencing and assembly procedures, one previously Sanger-sequenced BAC (Cm13_J04, Acc. No. EF657230.1) was selected from the MRGH63 contig constructed on the basis of BAC end information [12,35]. We also added to the pool BACs Cm43_H20 and Cm14_M22 of this contig that are known to overlap with the former (Additional file 1 Figure S1). In all, this first pool of BAC clones consists of 35 BACs mapping to 33 different loci.

A second batch of 20 BACs anchored to genetic markers distributed throughout the genome but different from those corresponding to the first set of 35 BACs was also chosen for 454-sequencing (see Figure 1). Three additional BACs were included in this second pool: the above-mentioned BAC Cm43_H20, and two randomly chosen BAC clones not linked to any known genetic marker (BACs Cm21_I02 and Cm12_I23). In all, the second pool consists of 23 BACs mapping to at least 21 different genetic loci.

In all, the selected two sets of BACs represent an estimated 7.5 Mb of the melon genome, based on BAC library average insert size. The complete list of selected BAC clones, together with their corresponding genetic markers can be found in Table 1. Information regarding genetic map position, marker type and references of the genetic markers can be found in the Additional File 2 Table S1.

Sequencing and assembly

Both shotgun and 3 kb paired-end libraries were constructed for each pool of BACs and the sequencing was carried out independently as described in the Methods section. A summary with the details of the different 454 runs, including number of reads, total length and average read size can be found in Table 2. In total, over one million reads representing 274 Mb of sequence from the 35 BACs pool and over 400,000 reads totaling 105 Mb from the 23 BACs pool were produced. The raw data (sff files) have been deposited in the SRA archive of the NCBI under the accession number SRA024701.1.

A global assembly of all reads from both BAC pools was performed as described in the Methods section. In addition, two independent assemblies were performed using reads from each pool. The reduced complexity in the separate assemblies of individual pools of BACs would suggest a more accurate assembly. Indeed, the number of contigs slightly increases and their size decreases in the global assembly, but overall, the result of the global assembly resembles the results from the assemblies of the individual BAC pools, except for a few cases. For example, in the case of the BAC Cm54_I13, we obtained a single scaffold in the 35 BACs pool assembly corresponding to two scaffolds from the global assembly. What separates the two scaffolds (when aligned to the single one) is a 273 bp gap flanked by several TA motives. On the other hand, scaffold00040 from the global assembly contained 631 additional nucleotides and a 522 bp long gap flanked by AT repeats at one of its extremes compared to its counterpart scaffold from the 23 BACs pool assembly. As we do not have a reference genome, we considered the larger scaffold as reference. A detailed summary of the whole process with the metrics of the three assemblies can be found in Table 3. Based on this information, we conclude that for assembling a modest number of BACs it is not worth separating them in smaller pools (increasing the sequencing costs), and if reduction of complexity is imperative (when dealing with very repetitive genomes, for example) then the extreme approach could be considered and barcode each BAC.

The assignment of contigs and scaffolds to BACs was performed using anchored genetic markers and BAC-end sequences as described in the Methods section. Also, the information from the *C. melo* FPC physical map [26] together with BAC-end sequences from some BAC clones in FPC contigs allowed us to manually edit

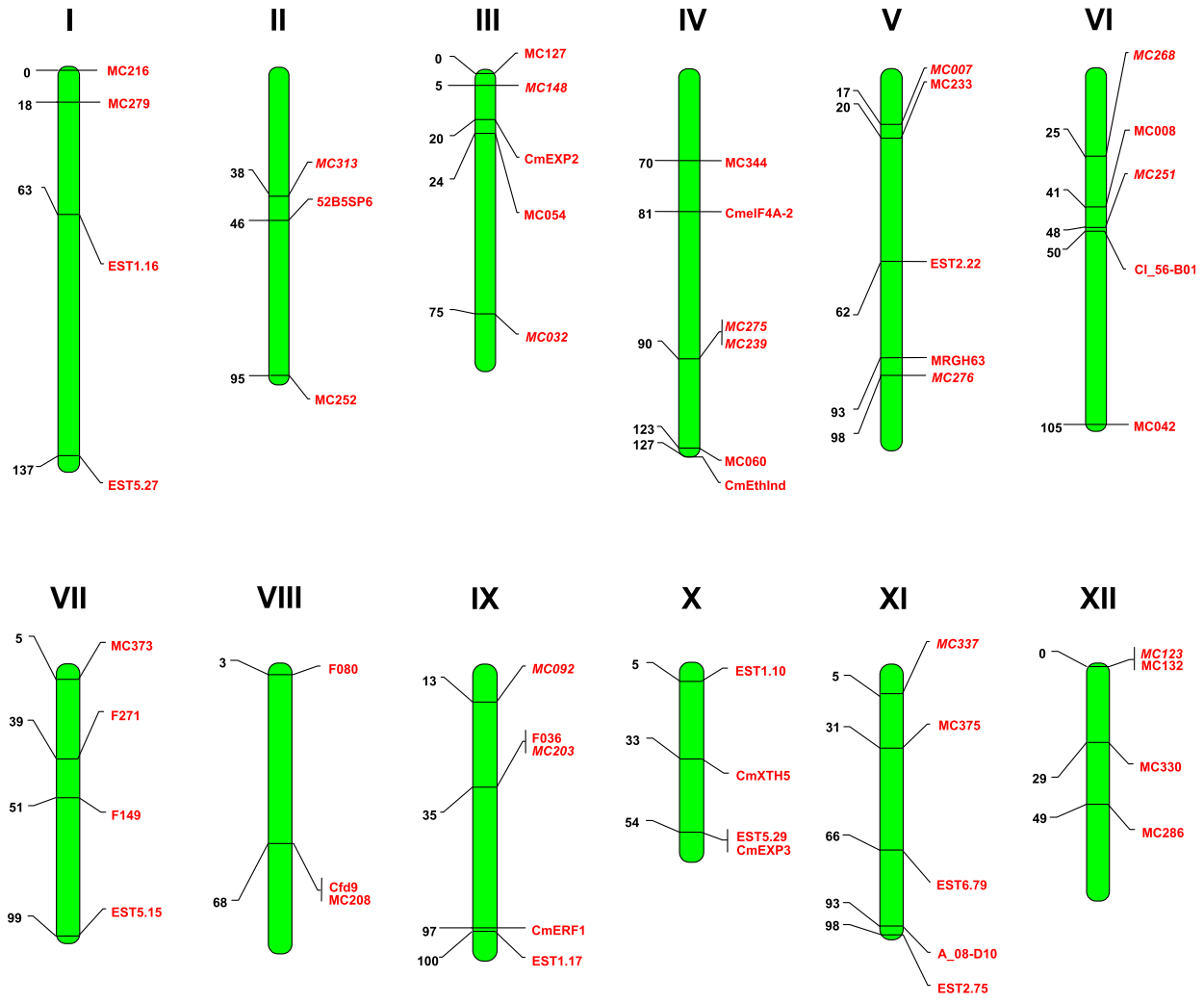


Figure 1 Distribution of the genetic markers used to anchor the sequenced BACs to the *C. melo* genetic map. Linkage groups are numbered according to the *C. melo* map of Deleu *et al.* [24]. Map distances are indicated on the left in cM. Markers in italics have been placed in an approximate position from Oliver *et al.* [21].

two scaffolds of the final assembly. The physical map was also useful in assigning BACs Cm21_I08 and Cm12_I23 to their corresponding scaffolds, as no genetic markers correspond to these BACs. Finally, the previously Sanger-sequenced BAC Cm60_K17 (Acc. No.: AF499727.1, [12]) was added to the alignment of the sequenced BACs from the MRGH63 contig in order to extend the sequence used for subsequent analysis (see Additional file 1 Figure S1).

The final assembly consists of 73 scaffolds totaling 6.3 Mb, 73% of which are longer than 60 kb, with average scaffold size 86.8 kb and the largest scaffold 304 kb long; also, 744 unscaffolded contigs totaling 382 kb of sequence remain (Table 3). The sequence coverage of the final assembly is 39x, calculated as the ratio between the total length of the sequence reads and the assembly sequence length. Paired-end reads are used in the process of sequence assembly to join contigs (formed by read alignments) in structures called scaffolds, which represent sorted and correctly orientated contigs that are separated by gaps which sizes are estimated based on the average paired-end size (see, for example, [39]). The N50 contig size of our assembly was rather small (30.6 kb) compared to the N50 scaffold size (107.6 kb). This result confirms the importance of paired-ends when it comes to assembling a complex genome using 454 sequences.

Regarding the assignment of sequences to particular BACs, BAC Cm47_C02 could not be assigned to any scaffold or contig and BAC Cm46_I24 was assigned to a small contig of less than 1 kb using the genetic marker sequence information, and to another two small scaffolds

Table 1 Correspondence between sequenced BAC clones, genetic markers and assembled contigs/scaffolds*

Linkage group	Marker name[a]	BAC name	Scaffold					
			Name	GenBank ID	Length (bp)	Stretches of Ns		BAC-ends Found[b]
						No.	Length (bp)	
I	MC216	Cm57_M11[2]	Contig311	HM854822	626	0	0	0
I	MC279	Cm31_J02[1]	Scaffold00087	HM854819	126,619	3	1,334	2
I	EST1.16	Cm33_F23[2]	Scaffold00078	HM854813	113,787	11	9,016	2
I	EST5.27	Cm43_O21[1]	Scaffold00052a	HM854797	131,697	10	3,452	2
II	MC313	Cm05_B01[2]	Scaffold0006	HM854766	126,054	4	4,978	2
II	52B5SP6	Cm52_B05[1]	Scaffold52B05	HM854760	138,922	38	29,090	2
II	MC252	Cm46_G13[1]	Scaffold0009	HM854768	151,031	2	529	2
III	MC127	Cm05_P10[2]	Scaffold05P10	HM854751	114,263	7	5,100	1
III	MC148	Cm45_K10[2]	Scaffold00035	HM854788	105,652	2	630	2
III	CmEXP2	Cm24_H21[1]	Scaffold0003[3]	HM854763	86,310	3	1,125	2
III	MC054	Cm52_C09[1]	Scaffold00024b	HM854780	61,053	2	666	1
III	MC032	Cm55_F19[1]	Scaffold55F19	HM854755	110,853	11	9,792	2
IV	MC344	Cm33_M05[2]	Scaffold00077	HM854812	148,622	6	2,636	2
IV	CmeIF4A-2	Cm59_B11[1]	Scaffold59B11	HM854756	100,000	5	5,558	2
IV	MC275	Cm11_I12[1]	Scaffold11I12	HM854757	110,000	9	6,210	2
IV	MC239	Cm06_A03[1]	Scaffold00012a	HM854770	75,205	0	0	1
IV	MC060	Cm46_O06[2]	Scaffold00041[3]	HM854791	103,741	2	668	2
IV	CmEthInd	Cm14_C18[1]	Scaffold0001	HM854762	108,322	1	478	2
V	MC007	Cm52_M23[2]	Scaffold00070	HM854810	112,968	2	567	2
V	MC233	Cm24_G05[1]	Scaffold00017	HM854775	82,645	1	247	2
			Contig00219	HM854821	810	0	0	0
V	EST2.22	Cm46_I24[1]	Scaffold00044	HM854793	12,974	4	4,832	1
			Scaffold00071	HM854811	20,426	7	9,835	1
V	MRGH63	Contig MRGH63:	ScaffoldMRGH63	HM854749	302,015	9	4,457	
		Cm13_J04[1,4]						2
		Cm14_M22[1]						2
		Cm43_H20[1,2]						2
V	MC276	Cm01_N3[1]	Scaffold00015	HM854773	180,444	5	2,607	1
VI	MC268	Cm02_C04[2]	Scaffold00031	HM854785	105,693	4	1,877	2
VI	MC008	Cm31_G08[2]	Scaffold00033	HM854786	109,145	3	764	2
VI	MC251	Cm02_K14[1]	Scaffold00058	HM854801	121,212	8	2,788	2
VI	CI_56-B01	Cm27_F03[1]	Scaffold27F03	HM854758	96,265	1	506	2
VI	MC042	Cm20_H14[1]	Scaffold00018	HM854776	96,294	3	891	2
VII	MC373	Cm55_C15[1]	Scaffold00057	HM854800	98,578	1	316	2
VII	F271	Cm45_K01[1]	Scaffold45K01	HM854759	100,000	11	7,803	2
VII	F149	Cm47_C02[2]	-	-	-	-	-	-
VII	EST5.15	Cm47_A05[1]	Scaffold0004	HM854764	101,589	3	1,515	2
VIII	F080	Cm22_K19[1]	Scaffold00081	HM854815	99,638	6	2,137	2
VIII	Cfd9	Cm06_D16[1]	Scaffold00025	HM854781	102,876	13	10,731	1
VIII	MC208	Cm19_K17[2]	Scaffold00023	HM854779	125,428	1	242	2
IX	MC092	Cm24_H03[2]	Scaffold24H03	HM854823	106,131	3	1,282	2
IX	F036	Cm34_G20[1]	Scaffold00069	HM854809	125,129	2	1,825	2
IX	MC203	Cm54_J04[2]	Scaffold54J04	HM854753	100,000	6	13,775	2
IX	CmERF1	Cm54_I13[1]	Scaffold54I13	HM854824	94,153	3	1,036	2
IX	EST1.17	Cm10_D04[1]	Scaffold10D04	HM854761	154,039	45	28,530	2
X	EST1.10	Cm03_A21[1]	Scaffold00079	HM854814	126,557	2	482	2
X	CmXTH5	Cm41_H09[1]	Scaffold0005	HM854765	136,275	2	696	2

Table 1 Correspondence between sequenced BAC clones, genetic markers and assembled contigs/scaffolds* *(Continued)*

X	EST5.29	Cm19_G01[2]	Scaffold00013	HM854771	100,283	14	8,175	2
X	CmEXP3	Cm54_E01[2]	Scaffold54E01	HM854754	100,000	18	13,832	2
XI	MC337	Cm12_F09[2]	Scaffold00028	HM854783	118,830	3	1,840	2
XI	MC375	Cm03_C12[1]	Scaffold00014	HM854772	128,906	13	9,538	1
XI	EST6.79	Cm59_N09[1]	Scaffold00085	HM854817	102,799	1	338	1
XI	A_08-D10	Cm24_I03[2]	Scaffold24I03	HM854752	121,276	16	8,734	1
XI	EST2.75	Cm33_O17[1]	Scaffold00051	HM854796	123,309	3	1,296	2
XII	MC123	Cm59_C10[2]	Scaffold59C10	HM854750	10,343	3	1,861	0
XII	MC132	Cm03_I02[1]	Scaffold00086	HM854818	79,495	4	3,249	1
XII	MC330	Cm09_A17[1]	Scaffold00034	HM854787	96,336	1	271	2
XII	MC286	Cm05_O10[2]	Scaffold00020	HM854778	142,670	8	2,708	2
-	-	Cm21_I08[2]	Scaffold00061	HM854803	146,020	12	5,169	2
-	-	Cm12_I23[2]	Scaffold00010	HM854769	114,336	9	8,517	2

*Additional information regarding sequence and annotation characteristics of the assembled sequence can be found in the Additional file 3 Table S2

[a]Genetic marker information can be found in the Additional file 2 Table S1

[b]One (1), both (2) or none (0) BAC-ends found on the scaffold/contig sequence

[1]First pool of BACs

[2]Second pool of BACs

[3]Marker sequence not found. Scaffold assignment based on information derived from the *C. melo* physical maphttp://melonomics.upv.es/static/files/public/physical_map/ and BAC-end information

[4]Sequenced previously by Shotgun-Sanger [35], Acc. No. EF657230

using both BAC-end sequences. All other BACs were assigned to a unique scaffold or contig, two of which were smaller than 15 kb, another five in the 60-90 kb range while the rest was over 90 kb long (Table 1).

The search for BAC ends in the final set of contigs and scaffolds suggests that at least 42 scaffolds cover the complete sequence of 44 BACs (including the three BACs belonging to the scaffold MRGH63). An average of seven stretches of Ns (produced as a result of contig scaffolding) was found per scaffold and the total length of all Ns accounts for 4.8% of the final assembly length (see Additional file 3 Table S2). Nine additional scaffolds assigned to as many BAC clones were found to contain only one BAC border each; however, six of these scaffolds were bigger than 100 kb, and so they probably represent complete BAC sequences but for small deletions at their borders, while the rest measured between 60 and 80 kb and could represent a significant proportion of their correspondent BAC sequences. Finally, BAC borders were absent from two BAC sequences (corresponding to BACs Cm57_M11 and Cm59_C10), both smaller than 11 kb and therefore most likely incomplete.

As a summary, of a total of 57 pooled BACs, most likely complete sequences were produced for 50 BAC clones, three were incomplete but in the range of 60-80 kb and four BACs were attributed very limited sequence information. As the assignment was performed using a small amount of sequence information, namely the marker and BAC-end sequences (not available for all BACs), any sequence shorter than the full BAC insert size has few chances of being assigned to

Table 2 Details of the 454 FLX runs from which sequence data were obtained

Pool	Sequencing plate regions	Library type	No. of reads	No. of Paired-end reads	Total length (bp)	Average read size (bp)
35 BACs						
	2/2[a]	Shotgun	445,232	-	110,498,601	248
	2/4	Paired end	89,392	3,152	23,214,413	260
	2/2	Paired end	557,452	126,681	139,772,537	251
23 BACs						
	2/2	Shotgun	261,304	-	64,679,158	247
	3/8	Paired end	155,166	56,990	40,110,640	259

[a]Includes 8,046 reads obtained from the titration process of the samples as well as 20,627 reads from a 1/2 region that was poorly sequenced

Table 3 Metrics for BAC assemblies and final results after manual correction.*

	35 BACs	23 BACs	Global assembly 57 BACs (two pools together)	Manual correction
No. of contigs[a]	514	247	797	-
No. of bases in contigs	3,936,343	2,325,066	6,127,262	-
Average contig size (bp)	7,658	9,413	7,687	-
N50 contig size (bp)	32,583	32,458	30,630	-
Largest contig size (bp)	117,242	112,451	123,360	-
Q40 plus bases	99.5%	99.5%	99.5%	-
No. of scaffolds	58	32	87	73
No. of scaffolds larger than 20 kb	41	25	62	57
No. of bases in scaffolds	4,040,161	2,307,575	6,206,490	6,340,685
Average scaffold size	69,657	72,111	71,338	86,882
N50 scaffold size	107,196	113,599	107,604	113,787
Largest scaffold size	222,620	200,453	212,424	303,725[b]
No. of unscaffolded contigs[c]	479	234	798	744
No. of bases in unscaff. contigs	224,871	121,734	417,982	382,726
Average unscaff. contig size	469	520	524	514
Coverage	x46	x25	x39	x39

*Reads from all 57 BACs were processed together in one assembly run. Additional assemblies of each BAC pool were independently done and served for comparison purposes and to manually correct some scaffolds in the global assembly

[a]Only contigs larger than 500 bp

[b]Two previously published BACs were included in this scaffold (see Methods section and Additional file 1 Figure S1)

[c]Contigs larger than 100 bp

any particular BAC. This is obvious for the BAC Cm46_I24 where with each BAC-end sequence and the marker sequence we assign three rather small sequences (Additional file 3 Table S2). In our results, a total of 374 kb distributed in 20 contigs/scaffolds longer than 2,000 bp remained unassigned after the final assembly and could account for most of the sequence of those four problematic BACs. All markers but one (F149), and all available BAC-end sequences but three, matched against a contig or scaffold. The nucleotide sequences of contigs and scaffolds assigned to BACs as well as of those unassigned assembly sequences larger than 2 kb have been deposited in the GenBank database and their accession numbers can be found in the Additional file 3 Table S2.

The number of gaps per Mb (61) and the estimated amount of missed sequence in our main assembly (5%) are lower than those values from the above-mentioned studies using 454 sequencing of BAC clones [37-39], a fact most probably due to the absence of paired-end sequencing in [38,40], to the short reads that were being produced at the erlier stages of 454 technology (100 bp on average) [38], to the complexity of the barley and salmon genomes as compared with melon's [38-40], and to the higher amount of assembled sequence in the case of *O. barthii* [37]. In summary, although using shotgun and paired-end libraries of pooled BACs remains a costly proposition for sequencing a whole genome, it is well adapted to certain situations. Indeed, our results show that it is a feasible and cost-efficient strategy for sequencing particular regions of interest of relatively compact genomes like that of melon. This approach would also be useful in genome walking strategies for gene cloning, or resolving a particular region where a physical map is available.

Sequence accuracy assessment

The quality of the final assembly was assessed by comparing the sequence from scaffold MRGH63 corresponding to BAC Cm13_J04 (Additional file 1 Figure S1) against the 99 kb-long sequence of the same BAC previously obtained using a shotgun-Sanger approach [35]. Table 4 shows the differences between the Sanger and 454 sequences. Apart from five small stretches of Ns representing 3.6% of the BAC length, the only other discrepancies are 15 homopolymeric regions and two dinucleotide tandem repeats. The differences in homopolymeric regions were found in 15 of the 26 mononucleotide repeats longer than 11 nt, and in all cases but one the 454 repeat resulted to be one to three nucleotides shorter than the Sanger sequence. It is interesting to note that no differences were found in the 896 mononucleotide repeats shorter than 11 nt. The discrepancies in dinucleotide tandem repeats affected two $(CT)_{15}$ and $(GA)_{21}$ microsatellites.

It has been already described that Sanger and 454 technologies have a generally comparable level of accuracy regarding genic regions or other single-copy

Table 4 Differences between Sanger- and 454-sequences of BAC Cm13_J04.

Length of Sanger-sequence	98,716 bp			
Stretches of Ns on 454-sequence	5	3,572 bp (3.6%)		
Homopolymers	**length**	**Sanger[1]**	**454 differences[2]**	
			No.	**Motif**
A/T				
	≤10	847	0	
	11	5	0	
	12	5	3	$(A/T)_{11}$
	13	3	2	$(A/T)_{12}$
	14	2	1	$(A/T)_{13}$
	15	3	2	$(A/T)_{14}$
			1	$(A/T)_{13}$
	16	1	1	$(A/T)_{14}$
	17	3	2	$(A/T)_{15}$
	18	1	1	$(A/T)_{17}$
	22	1	1	$(A/T)_{19}$
	24	1	0	
	28	1	1	$A_{15}CA_{13}$
C/G				
	5-7	49	0	
Other	**Sanger**	**454**		
	$(CT)_{15}$	$(CT)_{15}$CTACTTACTTACTTACNNNNNNNC$(CT)_{14}$		
	$(GA)_{21}$	$(GA)_{21}$GTAGTACGTACN$_{23}(GA)_6$		

[1]Number of homopolymers in the Sanger sequence

[2]Number of homopolymers in the 454 sequence showing differences with the corresponding homopolymers in the Sanger sequence

sequences, homopolymeric stretches being the main source of read errors in both techniques when low copy regions are considered [37,38,43,44]. Previous reports have also shown that longer stretches of A and T are more likely to cause problem when using pyrosequencing [38]. Indeed, there is a tendency of homopolymers to be shorter in the 454 sequence than in the Sanger reads, although at least a report exists where the stretches were consistently found to be one nucleotide longer in the 454 sequences [38,43]. In summary, the polymorphisms detected between the melon 454 and Sanger sequences in a 100 kb interval involved only 1.7 differences every 10,000 bp, a figure close to previously reported values [37,38].

Besides homopolymers, repetitive DNA is known to be more problematic for 454 sequencing than for Sanger due to the shorter length of the 454 reads. Repetitive regions can be collapsed into one consensus contig causing gaps to appear in the final assembly. This may be the main reason behind the gaps accounting for an estimated loss of *ca*. 5% of melon sequence in our final assembly. Indeed, all five stretches of Ns in Cm13_J04 consensus sequence are found in two regions that contain repetitive sequences such as a transposable element and a TIR-NBS-LRR resistance gene (data not shown).

Sequence annotation

Ab initio prediction of protein coding, tRNA and rRNA genes was carried out as described in the Methods chapter. The predictions were validated by homology with protein sequences at NCBI databases and with ESTs from the melon unigene v3 database at ICUGI [11]. A census of simple sequence repeats (SSRs) was also performed using the msatcommander software.

A summary of the sequence and annotation features of all 58 contigs and scaffolds longer than 20 kb, representing 6.2 Mb of genomic sequence, can be found in Table 5. As a whole, 616 protein coding genes (excluding transposons) were predicted, of which 73.2% were found to show homology with known *C. melo* ESTs. The average gene density is estimated to be 9.9 genes for each 100 kb but varies on the 2-20 range when individual scaffolds are considered; the average intron and exon length are respectively 393 bp and 238 bp and number of exons per gene is 4.9, with 46% of coding sequence being introns. Predicted proteins were 386 aa long on average. Regarding SSRs, 4,430 microsatellites were found representing 1.25% of the total sequence, about one SSR every 1.3 kb. The GC content composition was 33%, eleven tRNA genes were found in five BAC clones and no rRNA genes could be found in the analyzed sequence. Additional file 3 Table S2 contains a

more detailed report of the individual characteristics of each scaffold or contig larger than 2 kb.

The recent publication of the *Cucumber sativus* genome sequence begs the comparison of sequence and annotation characteristics of both cucurbit species [7]. Overall, the statistics of protein-coding genes from both cucurbits are quite similar. The predictions for the cucumber genome are a gene density of 10 per 100 kb, mean protein length of 349 amino acids, average number of exons per gene, exon length and intron length of 4.8, 238 bp and 483 bp, respectively, and tRNA gene density of 2.9 per Mb. While the gene density, mean exon length and average number of exons per gene are very similar in both species, in cucumber the protein length is only slightly smaller (0.9x), and mean intron length is just 1.2 times greater.

The apparent similar gene density, together with the similarity in average protein length, number of exons and average exon and intron lengths, seems contradictory with the difference in genome size between both species. Indeed, the estimated size of the melon genome is 1.3x that of cucumber [7,9]. It has to be taken into account, however, that the cucumber gene density was calculated based on as much as 70% of the complete genomic sequence, which most probably included gene-poor regions, while the melon gene density has been estimated using BAC clones that have gene- or EST-based genetic markers and thus probably represent gene-rich regions. Therefore, it might be the case that the actual melon gene density is lower than that of cucumber, hypothesis that is supported by the analysis of syntenic regions from both genomes (see below in the "Analysis of microsynteny" section).

Transposon content of the sequenced BACs

Transposons were annotated using sequence similarity searches with previously characterized transposons as well as by *Ab initio* methods based on transposon structural characteristics. As expected, most of the elements found belong to the retrotransposon class of mobile elements, with the *Gypsy* family being the most represented. However, the fraction of the genome these elements occupy is apparently smaller than in other genomes of similar size. Indeed, while retrotransposons account for 20% of the genomes of grapevine (504.6 Mb) and *Lotus japonicus* (472 Mb) [45,46], these elements seem to account for only 7.2% of the melon genome (454 Mb) (Table. 6). Retrotransposons are not randomly distributed in genomes and while some elements preferentially integrate in gene-rich regions (see for example [47]), others target heterochromatic regions for integration, in particular those belonging to the *Gypsy* family which are usually present at higher copy number [48]. Thus, the apparent low retrotransposon copy number could be due to the fact that heterochromatic regions are under-represented in the 1.5% fraction of the genome analyzed, which was selected to be representative of the gene-rich regions of the melon genome.

We have also found representatives of all the major families of DNA transposons, including CACTA, MULE,

Table 5 *C. melo* BAC sequences characteristics[a]

Total sequence length	6,230,040 bp
Sequence length excluding stretches of Ns	5,958,994 bp
Number of predicted protein coding genes[b]	616
Number of predicted protein coding genes with homology to *C. melo* ESTs	451 (73.2%)
tRNA genes	11
Gene density[c]	9.9 genes/100 kb (1.5 - 19.7, SD: 4.3)
Average exon length	238 bp
Average intron length	393 bp
Exons per gene	4.9 (1-29, SD: 4.4) (74% of genes ≤ 6 exons) (23% intronless)
Average protein length[d]	386 (34-2,156, SD: 268)
Average% of introns in coding sequence[e]	45.6 (4.3 - 95.5, SD: 20.6)
GC content (%)	33 (30.2 - 38.7, SD: 1.34)
SSRs[f]	4,430 (74,590 bp, 1.25% of total sequence) 1 SSRs/1.3 Kb
Transposable elements[g]	139

[a]From the analysis of all 57 scaffolds plus one contig longer than 20 kb

[b]Genes from transposons not counted

[c]Partial genes at BAC borders counted as 0.5 genes

[d]Transposon proteins not considered

[e] ORFs without introns not considered

[f]Minimum repeat lengths considered: 10 bp (mononuc.), 12 bp (di- and trinuc.), 16 bp (tetranuc.), 20 bp (pentanuc.) and 24 bp (hexanuc.)

[g]See Table 6 for a more detailed analysis of transposon content

hAT, PIF and Helitron elements, covering in total 0.93% of the analyzed sequence (Table 6), which is consistent with what has been reported for the genomes of grapevine (1.98%) [49] and *Lotus japonicus* (0.97%) [46].

Analysis of microsynteny

Four of the longest scaffolds (9, 15, 77 and MRGH63, totalling 782 kb) were used to search the cucumber genome assembly [50] for syntenic regions, as described in the Methods section. As it can be expected from the close phylogenetic relatedness of these two species, a high degree of collinearity was found in all four regions analysed (Figure 2). The relative syntenic quality (see the Methods section) ranged from 84% (for scaffold MRGH63) to 97% (for scaffold00015), averaging 92%, and the homologous protein sequences rendered in all cases e-values lower than 1E-46 with an average identity of 87% when aligned using BLASTP (see Additional file 4 Table S3). Regarding the annotation characteristics of the predicted genes, the average protein lengths of the four melon regions analyzed were x0.8-x1.2 those of cucumber, with the syntenic melon genes being, as an average, only x0.96 smaller than the cucumber ones; the average number of exons of the melon syntenic regions were x0.84-x1.1 those of the cucumber regions, with the syntenic melon genes having, as an average, only x0.92 less exons than the cucumber ones; also, although the average exon length of all syntenic melon genes was almost identical to that of cucumber, the average intron length of the syntenic melon genes was x1.3 that of their cucumber counterparts (Additional file 4 Table S3).

Besides, the orientation of the putative syntenic genes was found to be conserved in all cases. However, a number of genes were duplicated in melon. These included the expansion of a cluster of NBS-LRR genes present in scaffold MRGH63, which is particularly interesting as the *Vat* gene and other disease resistance genes have been mapped to this region [33]. NBS-LRR genes are the main family of resistance genes in plants, and are frequently found in clusters [51]. Highly conserved gene order and content together with 95% of sequence similarity over coding regions has already been reported by Huang *et al*. based on the comparison of four sequenced BAC clones against the sequenced cucumber genome [7].

Besides the duplication of several genes, a major difference between cucumber and melon syntenic regions is the number of transposon insertions [Figure 2]. The cucumber sequences analysed contain only two retrotransposon insertions, one of which seems very old as it is highly degenerated. On the contrary, the melon syntenic regions contain three DNA transposons (two hATs and one MULE) and 15 retrotransposons (most of them from the *Gypsy* superfamily), including the degenerated retrotransposon found in cucumber. In particular, transposon activity appears to account for the expansion of *ca*. 60 kb in the melon scaffold0077 relative to its cucumber counterpart. In scaffold MRGH63, a localised transposon number amplification together with duplication of melon resistance gene homologs (see below) accounts for an 88 kb-long expansion of the sequence of melon relative to that of cucumber. Also, transposons were found to be putatively involved in gene disruption processes in scaffolds 9 and MRGH63.

These results suggest that transposition activity after the divergence of the two ancestors of melon and cucumber has been low in cucumber but very high in melon. This transposon amplification and mobilization could be a reason for the 1.8× increase in size of the melon syntenic regions. Bearing in mind that the melon genome is estimated to be 1.3× greater that of cucumber, it can tentatively be assumed that transposon activity may be mainly responsible for that difference in genome sizes.

It is interesting to note that almost half of the melon specific transposons are interspersed with NBS-LRR predicted genes that potentially form resistance gene clusters. Gene duplications and transposon insertions have been proposed to provide a structural environment that permits unequal crossovers and interlocus gene conversion allowing rapid evolution of resistance genes [51]. In addition, the presence of active retrotransposons interspersed with resistance genes may also contribute to the resistance gene regulation by silencing related mechanisms [52]. A detailed analysis of syntenic regions containing putative resistance genes between melon and cucumber may provide new information on the

Table 6 Transposon content in the *C. melo* sequenced BACs[a]

Family	Copies (no.)	Total lenght (bp)	Analyzed sequence (%)
DNA transposons			
CACTA	15	30,238	0.48
hAT	4	8,726	0.14
MULE	6	17,836	0.28
PIF	1	842	0.01
helitron	1	830	0.01
Total	**27**	**58,472**	**0.93**
Retrotransposons			
Copia	15	49,606	0.79
Gypsy	18	80,452	1.28
Non-LTR	3	9,664	0.15
Non-classified	77	313,326	5.0
Total	**113**	**453,048**	**7.2**

[a]From the analysis of all contigs and scaffolds longer than 2 kb

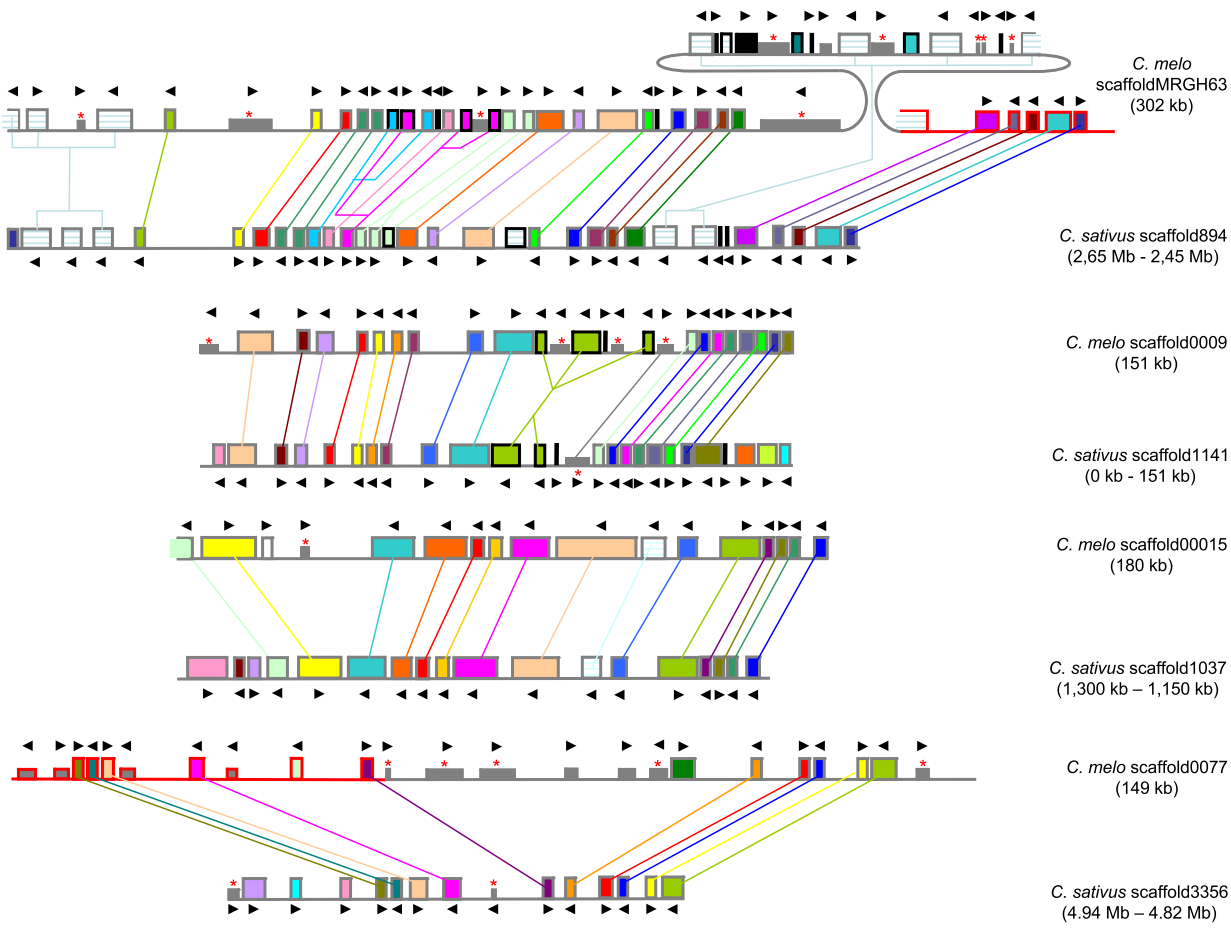


Figure 2 Overview of microsynteny between four melon scaffolds and four regions in the *C. sativus* genome. Genes are represented by square blocks. Homologous genes are illustrated with the same colour and indicated by connecting lines of the corresponding colour. *Ab initio* predicted genes with no homology to public EST or protein databases are shown in black. Transposable elements are in gray, with red asterisks as an additional mark for retrotransposons. Genes coding for NBS-LRR disease resistance proteins are represented by square block filled with blue vertical lines. Putative pseudogenes are depicted as black edge boxes. The annotation of *C. melo* scaffoldMRGH63 and scaffold00077 was complemented using information from *ca.* 57 kb and 96 kb, respectively, of unpublished melon sequence (represented in the figure as red edge boxes). Figure drawn to scale.

evolution of resistance genes and the development of new resistances in cultivated crops.

Conclusion

A set of 57 BAC clones from a double haploid line of melon was sequenced in two pools with the 454 system using both shotgun and paired-end approaches followed by bioinformatic assembly of the fragments obtained. From this assembly it was possible to obtain most likely complete sequences for 50 of these BACs, as judged by the length and the presence of BAC-end sequences, with a final coverage of 39×. The accuracy of the assembly was excellent, compared with a BAC clone already sequenced with the Sanger method, except in a small number of repetitive sequences, consistent with other 454 sequencing projects [37,38]. These results show that 454-sequencing of pooled BACs, using both shotgun and paired-end libraries, is a feasible strategy for sequencing long stretches of genomic sequence from medium-size genomes such as that of melon. However, correction using other sequencing techniques would be needed for medium to high repetitive content regions.

The analysis of the fraction (around 1.5%) of the melon genome obtained provides a pilot overview of this species' genomic structure. Predicted gene annotations were confirmed in 73% of the cases by comparison with EST collections. This is probably a good measure of the completeness of the transcriptome information currently available for this species. The analysis of the sequences provides an interesting overview of the features such as microsatellite content, gene density and average protein length, revealing similarity to that of its close relative, cucumber.

Finally, the comparison of four melon regions totalling 782 kb against the genomic sequence of cucumber (the only other Cucurbit species where a draft genome sequence is available) reveals a high degree of collinearity between both species. The analysis of the detected syntenic regions suggests that the size difference of the two genomes is due to the expansion of intergenic regions, mainly through the activity of transposable elements in melon after the divergence of the two species. It is particularly interesting to note that almost half of the detected melon-specific transposons are interspersed with NBS-LRR predicted genes that potentially form resistance gene clusters. We have confirmed the utility of this sequencing method for small genomic fractions, and the analysis of the data thus obtained has expanded our understanding of the melon genome structure and the mechanisms underlying its evolution.

Methods

BAC library

A *Bam*HI BAC library from the double-haploid melon line 'PIT92' (PI 161375 × T111) was previously developed in our laboratory using pECBAC1 as cloning vector [[12], http://hbz7.tamu.edu/homelinks/bac_est/vector/sequence/sequence.htm]. With 23,040 BAC clones distributed in sixty 384-well plates, an average insert size of 139 kb and 20% empty clones, the library represents 5.7 genomic equivalents of the haploid melon genome.

DNA extraction

Two pools of 35 and 23 BACs were selected for the analysis. Individual preinocules were grown on 1 ml 1 × LB plus 12.5 μg/ml chloramphenicol at 300 rpm, 37°C, for 17 h. The following day, 30 μl of each BAC clone from the preinocules were added into 50 ml tubes containing 20 ml 1 × LB plus 12.5 μg/ml chloramphenicol, and grown at 37° C, 300 rpm for 15 h. The grown cultures were then mixed to produce two separate volumes representing the two BAC pools and the bacterial cells were harvested by centrifugation at 6,000 × g for 15 min at 4°C.

Genomic DNA-free BAC DNA extraction was performed using the QIAGEN® Large-Construct Kit (Cat. No. 12462) following the manufacturer's instructions. Both final DNA pellets were resuspended in 500 μl TE pH 8.0 each.

DNA sequencing

All sequencing was performed with a Roche 454 Genome Sequencer machine using FLX chemistry. Two DNA extractions were done from the 35-BACs pool, one to create a shotgun library and the other one to create a 3 kb paired-end library. The shotgun library was used for one titration run and one full run performed by Lifequencing S. L. at their premises in Valencia, Spain. The paired-end library was sequenced on two quarters of a plate followed by a full run performed at our 454 sequencing facility. For the 23-BACs pool, one DNA extraction was done which served to create a shotgun and a 3 kb paired-end library. The shotgun library was sequenced with a full run while the paired-end library was sequenced on three eighths of a plate; both runs were performed at our 454 sequencing facility.

Sequence assembly

Sequence assembly was done using Newbler version 2.3 with default parameters. Reads from all BACs were processed together in one assembly run. The sequence of *E. coli* strain DH10B (NC 010473.1) was used as screening database and the vector pECBAC1 as trimming database, but without 30 bp of sequence flanking either side of the BamHI restriction site (see below) http://hbz7.tamu.edu/homelinks/bac_est/vector/sequence/sequence.htm. Additional assemblies of each BAC pool were independently done using Newbler versions 2.3 and 2.0 (data now shown); results of these assemblies served for comparison purposes and only in some cases helped to manually correct some scaffolds in the global assembly.

Sequences of the genetic markers previously anchored to the analyzed BACs as well as some BAC-end sequences previously available in our laboratories (GenBank Acc. Nos. can be found in the Additional file 3 Table S2) were used to assign a sequence to a specific BAC. Based on this information, in some cases we could join two scaffolds that corresponded to the same BAC into a single superscaffold that would represent the whole BAC insert. In these cases a gap was introduced between the scaffolds so that the final sequence had the size of the average insert size of the BAC library. The manually introduced gaps accounted for 7.25% of all the gaps in the assembly. The sizes of these gaps in nucleotides are as follow: 500 in Scaffold52B05; 1,209 in Scaffold45K01; 1,538 in Scaffold11I12; 1,831 in Scaffold54E01; 2,288 in Scaffold55F19; 2,586 in Scaffold59B11; and 12,064 in Scaffold54J04.

In order to study how many of the assembled contigs and scaffolds represented the complete sequence of BACs, those sequences were searched for BAC borders in the following ways: 1) by searching at their extremes the 30 bp sequence corresponding to pECBAC1; 2) by blasting against individual reads containing the 30 bp sequence and 3) by blasting against BAC-end sequences that were already available for some of the sequenced BACs [see Additional File 3 Table S2].

Sequence annotation

Ab initio gene prediction was performed using the command-line version of Augustus 2.3 software http://augustus.gobics.de/ using *A. thaliana* as plant model.

The melon unigene v3 collection at ICUGI [11] was used to improve the Augustus prediction, setting the minimum identity parameter to 92. In some cases, the FGENESH annotation software at http://linux1.softberry.com/berry.phtml, with *Arabidopsis* as plant model, was used to complement or improve the Augustus annotation. The predicted coding sequences were checked against the non-redundant protein databases at NCBI using BLASTP searching for protein homologs.

tRNA genes were predicted using the tRNAscan-SE 1.21 software http://lowelab.ucsc.edu/tRNAscan-SE/ and rRNA genes were identified with RNAmmer 1.2 server http://www.cbs.dtu.dk/services/RNAmmer/. Simple sequence repeats (SSRs) were searched for using the msatcommander 0.8.2 software http://code.google.com/p/msatcommander/; the minimum repeat lengths considered were: 10 bp (mononucleotides), 12 bp (di- and trinucleotides), 16 bp (tetranucleotides), 20 bp (pentanucleotides) and 24 bp (hexanucleotides).

Transposons were annotated using *Ab initio* and sequence similarity searches integrated in a pipeline based on Dawgpaws [53]. The programs used for *de novo* prediction of LTR retrotransposons included LTR_STRUC [54], LTR_finder [55] and LTR_seq [56], and vary in the type of structures they look for, their stringency and their search algorithms. The homology-based approach consisted of searching for sequences that show a high degree of similarity to known transposons. For this, we compiled nucleotide databases of already characterized transposons obtained from the RepBase database [57] as well as NCBI [58]. Likewise, we constructed protein sequence databases of transposases from various transposon families, searching NCBI for combinations of keywords such as "transposase" and "CACTA", "hAT", "Mariner", "Mutator" or "PIF". This approach is useful for corroborating results obtained from the *de novo* programs, as well as identifying other types of transposons such as DNA transposons. The output of these various programs was converted into gff3 format and uploaded into the Apollo genome viewer and annotation tool [59], along with the gene annotations, for manual curation. As a first step, each scaffold was examined and putative transposons were identified according to the computational evidence. These were then manually inspected to look for LTRs or TIRs, query NCBI to determine which family they belong to, and resolve instances of nested or truncated elements. These *bona fide* transposons were used to query the set of scaffolds in similarity searches, aiming at identifying partial or degenerated copies and defining transposon families. This third step is particularly relevant when a large amount of sequence data is available, as aligning many copies of an element aids to precisely define its borders and find consensus sequences. At this point, with the current fraction of the genome available, we have not found enough copies of a single element to perform this part of the analysis.

Synteny analysis

Four annotated melon scaffolds were analysed for homology with the *Cucumis sativus* genome assembly deposited in Phytozome v5 [50], using the BLASTN algorithm. The selected cucumber regions were annotated the same way as the melon BACs. Pairs of homologous genes were tentatively selected on the basis of the gene annotation and then confirmed by performing BLASTP alignments of the correspondent predicted proteins. Syntenic regions were defined as contiguous regions containing two or more homologous genes in *C. melo* and *C. sativus*, irrespective of orientation and exact order of genes, based on the results of BLASTP comparisons. The relative syntenic quality in a region, expressed as a percentage, was calculated by dividing the sum of the conserved genes in both syntenic regions by the sum of the total number of genes in both regions, excluding transposable elements and collapsing tandem duplications [60].

Note

The C. melo BAC nucleotide sequences are available in the DDBJ/EMBL/GenBank databases under the accession numbers HM854749-HM854824. The raw data can be found in the SRA archive of the NCBI under the accession number SRA024701.1.

Additional material

Additional file 1: Figure S1. Schematic representation of the MRGH63 contig.

Additional file 2: Table S1. Genetic markers anchored to the sequenced BAC clones.

Additional file 3: Table S2. Sequence and annotation characteristics of the assembled scaffolds and contigs.

Additional file 4: Table S3. Annotation characteristics of the *C. melo* and *C. sativus* syntenic regions

Acknowledgements

We gratefully acknowledge Lifesequencing S. L. for technical assistance in 454-sequencing one of the DNA pools. This project was funded by the Plan Nacional de Investigación Científica of the Spanish Ministerio de Educación y Ciencia (Projects BIO2007-61789 to PPR and AGL2006-12780-C02-01 to JGM), by the Consolider-Ingenio 2010 Programme of the Spanish Ministerio de Ciencia e Innovación (CSD2007-00036 "Center for Research in Agrigenomics"), and by the Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya. We acknowledge the valuable technical help from Roche 454 and Roche Spain.

Author details

[1]Molecular Genetics Department, Center for Research in Agricultural Genomics CRAG (CSIC-IRTA-UAB), Jordi Girona, 18-26, 08034 Barcelona, Spain. [2]IRTA, Center for Research in Agricultural Genomics CRAG (CSIC-IRTA-UAB), Carretera de Cabrils Km 2, 08348 (Barcelona), Spain.

Authors' contributions

VMG conducted BAC DNA extractions, helped to manually correct the final sequence assembly, annotated scaffolds, and drafted the manuscript, AB led the pre-processing of the sequence raw data, produced the sequence assemblies, and helped drafting the manuscript, EMH and JMC were in charge of the transposon content analysis and helped drafting the manuscript, GM constructed the 23 BAC pool shotgun and all paired-end libraries and performed the sequencing reactions, JGM participated in the project design, coordinated the BAC sequencing, participated in the discussion of results, and helped to draft the manuscript, PP conceived and coordinated the project, and helped to draft the manuscript. All authors read and approved the final manuscript.

Received: 9 August 2010 Accepted: 12 November 2010
Published: 12 November 2010

References

1. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**:796-815.
2. The International Brachypodium Initiative: **Genome sequencing and analysis of the model grass *Brachypodium distachyon*.** *Nature* 2010, **463**:763-768.
3. International Rice Sequencing Project: **The map-based sequence of the rice genome.** *Nature* 2005, **436**:793-800.
4. Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du FY, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, *et al*: **B73 Maize Genome: Complexity, diversity, and dynamics.** *Science* 2009, **326**:1112-1115.
5. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagi M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman , Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS: **The Sorghum bicolour genome and the diversification of grasses.** *Nature* 2009, **457**:551-556.
6. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, *et al*: **Genome sequence of the palaeopolyploid soybean.** *Nature* 2010, **463**:178-183.
7. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, Ren Y, Zhu H, Li J, Lin K, Jin W, Fei Z, Li G, Staub J, Kilian A, van der Vossen EAG, Wu Y, Guo J, He J, Jia Z, Ren Y, Tian G, Lu Yao, Ruan J, Quian W, Wang M, Huang Q, Li B, Xuan Z, Cao J, Asan , Wu Z, Zhang J, Cai Q, Bai Y, Zhao B, Han Y, Ying Li, Li X, Wang S, Shi Q, Liu S, Cho WK, Kim JY, Xu Y, Heller-Uszynska K, Miao H, Cheng Z, Zhang S, Wu J, Yang Y, Kang H, Man Li, Liang H, Ren X, Shi Z, Wen M, Jian M, Yang H, Zhang G, Yang Z, Chen R, Liu S, Li J, Ma L, Liu H, Zhou Y, Zhao Y, Fang X, Li G, Fang Li, Li Y, Liu D, Zheng H, Zhang Y, Qin N, Li Z, Yang G, Yang S, Bolund L, Kristiansen K, Zheng H, Li S, Zhang X, Yang H, Wang J, Sun R, Zhang B, Jiang S, Wang J, Du Y, Li S: **The genome of the cucumber, *Cucumis sativus* L.** *Nature Genetics* 2009, **41**:1275-1283.
8. Arumuganathan K, Earle ED: **Nuclear DNA content of some important plant species.** *Plant Mol Biol Rep* 1991, **9**:208-218.
9. Ezura H, Fukino N: **Research tools for functional genomics in melon (*Cucumis melo* L.): Current status and prospects.** *Plant Biotechnology* 2009, **26**:359-368.
10. Gonzalez-Ibeas D, Blanca J, Roig C, González-To M, Picó B, Truniger V, Gómez P, Deleu W, Caño-Delgado A, Arús P, Nuez F, Garcia-Mas J, Puigdomènech P, Aranda MA: **MELOGEN: an EST database for melon functional genomics.** *BMC Genomics* 2007, **8**:306.
11. **The International Cucurbit Genomics Initiative (ICuGI).** [http://www.icugi.org].
12. van Leeuwen H, Monfort A, Zhang HB, Puigdomènech P: **Identification and characterization of a melon genomic region containing a resistance gene cluster from a constructed BAC library. Microlinearity between *Cucumis melo* and *Arabidopsis thaliana*.** *Plant Mol Biol* 2003, **51**:703-718.
13. Luo M, Wang YH, Frisch D, Joobeur T, Wing RA, Dean RA: **Melon bacterial artificial chromosome (BAC) library construction using improved methods and identification of clones linked to the locus conferring resistance to melon Fusarium wilt (*Fom-2*).** *Genome* 2001, **44**:154-162.
14. Mascarell-Creus A, Cañizares J, Vilarrasa-Blasi J, Mora-García S, Blanca J, Gonzalez-Ibeas D, Saladié M, Roig C, Deleu W, Picó-Silvent B, López-Bigas N, Aranda MA, Garcia-Mas J, Nuez F, Puigdomènech P, Caño-Delgado AI: **An oligo-based microarray offers novel transcriptomic approaches for the analysis of pathogen resistance and fruit quality traits in melon (*Cucumis melo* L.).** *BMC Genomics* 2009, **10**:467.
15. Ophir R, Eshed R, Harel-Beja R, Tzuri G, Portnoy V, Burger Y, Uliel S, Katzir N, Sherman A: **High-throughput marker discovery in melon using a self-designed oligo microarray.** *BMC Genomics* 2010, **11**:269.
16. Tadmor Y, Katzir N, Meir A, Yaniv-Yaakov A, Sa'ar U, Baumkoler F, Lavee T, Lewinsohn E, Schaffer A, Buerger J: **Induced mutagenesis to augment the natural genetic variability of melon (*Cucumis melo* L.).** *Israel J Plant Sci* 2007, **55**:159-169.
17. Nieto C, Piron F, Dalmais M, Marco CF, Moriones E, Gómez-Guillamón ML, Truniger V, Gómez P, Garcia-Mas J, Aranda MA, Bendahmane A: **EcoTILLING for the identification of alleclic variants of melon eIF4E, a factor that controls virus susceptibility.** *BMC Plant Biol* 2007, **7**:34.
18. Eduardo I, Arús P, Monforte AJ: **Development of a genomic library of near isogenic lines (NILs) in melon (*Cucumis melo L.*) from the exotic accession PI161375.** *Theor Appl Genet* 2005, **112**:139-148.
19. Wang YH, Thomas CE, Dean RA: **A genetic map of melon (*Cucumis melo* L.) based on amplified fragment length polymorphism (AFLP) markers.** *Theor Appl Genet* 1997, **95**:791-798.
20. Danin-Poleg Y, Reis N, Baudracco-Arnas S, Pitrat M, Staub JE, Oliver M, Arus P, deVicente CM, Katzir N: **Simple sequence repeats in *Cucumis* mapping and map merging.** *Genome* 2000, **43(6)**:963-974.
21. Oliver M, Garcia-Mas J, Cardus M, Pueyo N, Lopez-Sese AL, Arroyo M, Gomez-Paniagua H, Arus P, de Vicente MC: **Construction of a referente map of melon.** *Genome* 2001, **44**:836-845.
22. Silberstein L, Kovalski I, Brotman Y, Perin C, Dogimont C, Pitrat M, Klingler J, Thompson G, Portnoy V, Katzir N, Perl-Treves R: **Linkage map of *Cucumis melo* including phenotypic traits and sequence-characterized genes.** *Genome* 2003, **46**:761-773.
23. Gonzalo MJ, Oliver M, Garcia-Mas J, Monfort A, Dolcet-Sanjuan R, Katzir N, Arus P, Monforte AJ: **Simple-sequence repeat markers used in merging linkage maps of melon (*Cucumis melo* L.).** *Theor Appl Genet* 2005, **110**:802-811.
24. Deleu W, Esteras C, Roig C, González-To M, Fernández-Silva I, González-Ibeas D, Blanca J, Aranda MA, Arús P, Nuez F, Monforte AJ, Picó MB, Garcia-Mas J: **A set of EST-SNPs for map saturation and cultivar identification in melon.** *BMC Plant Biology* 2009, **9**:90.
25. Harel-Beja R, Tzuri G, Portnoy V, Lotan-Pompan M, Lev S, Cohen S, Dai N, Yeselson L, Meir A, Libhaber SE, Avisar E, Melame T, van Koert P, Verbakel H, Hofstede R, Volpin H, Oliver M, Fougedoire A, Stalh C, Fauve J, Copes B, Fei Z, Giovannoni J, Ori N, Lewinsohn E, Sherman A, Burger J, Tadmor Y, Schaffer AA, Katzir N: **A genetic map of melon highly enriched with fruit quality QTLs and EST markers, including sugar and carotenoid metabolism genes.** *Theor Appl Genet* 2010, **121(3)**:511-33.
26. González V, Garcia-Mas J, Arús P, Puigdomènech P: **Generation of a BAC-based physical map of the melon genome.** *BMC Genomics* 2010, **11**:339.
27. Nieto C, Morales M, Orjeda G, Clepet C, Monfort A, Sturbois B, Puigdomènech P, Pitrat M, Caboche M, Dogimont C, Garcia-Mas J, Aranda MA, Bendahmane A: **An eIF4E allele confers resistance to an uncapped and non-polyadenylated RNA virus in melon.** *Plant J* 2006, **48**:452-462.
28. Joobeur T, King JJ, Nolin SJ, Thomas CE, Dean RA: **The Fusarium wilt resistance locus Fom-2 of melon contains a single resistance gene with complex features.** *Plant Journal* 2004, **39**:283-297.
29. Boualem A, Mohamed F, Fernandez R, Troadec C, Martin A, Morin H, Sari MA, Collin F, Flowers JM, Pitrat M, Purugganan MD, Dogimont C, Bendahmane A: **A conserved mutation in an ethylene biosynthesis enzyme leads to andromonoecy in melons.** *Science* 2008, **321**:836-838.
30. Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, Pitrat M, Dogimont C, Bendahmane A: **A transposon-induced epigenetic change leads to sex determination in melon.** *Nature* 2009, **461**:1135-1138.
31. Moreno E, Obando JM, Dos-Santos N, Fernández-Trujillo JP, Monforte AJ, Jordi Garcia-Mas: **Candidate genes and QTLs for fruit ripening and softening in melon.** *Theor App Genet* 2008, **116(4)**:589-602.

32. Ezura H, Owino WO: **Melon, an alternative model plant for elucidating fruit ripening.** *Plant Science* 2008, **175**:121-129.
33. van Leeuwen H, Monfort A, Zhang HB, Puigdomènech P: **Identification and characterization of a melon genomic region containing a resistance gene cluster from a constructed BAC library. Microlinearity between *Cucumis melo* and *Arabidopsis thaliana*.** *Plant Mol Biol* 2003, **51**:703-718.
34. van Leeuwen H, Garcia-Mas J, Coca M, Puigdomènech P, Monfort A: **Analysis of the melon genome in regions encompassing TIR-NBS-LRR resistance genes.** *Mol Gen Genom* 2005, **273**:240-251.
35. Deleu W, González V, Monfort A, Bendahmane , Puigdomènech P, Arús P, Garcia-Mas J: **Structure of two melon regions reveals high microsynteny with sequenced plant species.** *Mol Genet Genomics* 2007, **278**:611-622.
36. Varshney RK, Nayak SN, May GD, Jackson SA: **Next-generation sequencing technologies and their implications to crop genetics and breeding.** *Trends in Biotechnology* 2009, **27**:522-530.
37. Rounsley S, Marri PR, Yu Y, He R, Sisneros N, Goicoechea JL, Lee SJ, Angelova A, Kudrna D, Luo M, Affourtit J, Desany B, Knight J, Niazi F, Egholm M, Wing RA: ***De novo* next generation sequencing of plant genomes.** *Rice* 2009, **2**:35-43.
38. Wicker T, Schlagenhauf , Graner A, Close TJ, Keller B, Stein N: **454 sequencing put to the test using the complex genome of barley.** *BMC Genomics* 2006, **7**:275.
39. Quinn NL, Levenkova N, Chow W, Bouffard P, Boroevich KA, Knight JR, Jarvie TP, Lubieniecki KP, Desany BA, Koop BF, Harkins TT, Davidson WS: **Assessing the feasibility of GS FLX pyrosequencing for sequencing the Atlantic salmon genome.** *BMC Genomics* 2008, **9**:404.
40. Steuernagel B, Taudien S, Gundlach H, Seidel M, Ariyadasa R, Schulte D, Petzold A, Felder M, Graner A, Scholz U, Mayer KFX, Platzer M, Stein N: ***De novo* 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley.** *BMC Genomics* 2009, **10**:547.
41. Morales M, roig E, Monforte AJ, Arús P, Garcia-Mas J: **Single-nucleotide polymorphisms detected in expressed sequence tags of melon (*Cucumis melo* L.).** *Genome* 2004, **47(2)**:352-60.
42. Essafi A, Diaz-Pendon JA, Moriones E, Monforte AJ, Garcia-Mas J, Martin-Hernandez AM: **Dissection of the oligogenic resistance to *Cucumber mosaic virus in the melon accession PI161375*.** *Theor Appl Genet* 2009, **118(2)**:275-284.
43. Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM, Soltis DE: **Rapid and accurate pyrosequencing of angiosperm plastid genomes.** *BMC Plant Biol* 2006, **6(1)**:17.
44. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437**:376-380.
45. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Demattè L, Mraz A, Battilana J, Stormo K, Costa F, Tao Q, Si-Ammour A, Harkins T, Lackey A, Perbost C, Taillon B, Stella A, Solovyev V, Fawcett JA, Sterck L, Vandepoele K, Grando SM, Toppo S, Moser C, Lanchbury J, Bogden R, Skolnick M, Sgaramella V, Bhatnagar SK, Fontana P, Gutin A, Van de Peer Y, Salamini F, Viola R: **A high quality draft consensus sequence of the genome of a heterozygous grapevine variety.** *PLoS One* 2007, **2(12)**:e1326.
46. Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, Fujishiro T, Katoh M, Kohara M, Kishida Y, Minami C, Nakayama S, Nakazaki N, Shimizu Y, Shinpo S, Takahashi C, Wada T, Yamada M, Ohmido N, Hayashi M, Fukui K, Baba T, Nakamichi T, Mori H, Tabata S: **Genome structure of the legume, Lotus japonicus.** *DNA Res* 2008, **15(4)**:227-39.
47. Le QH, Melayah D, Bonnivard E, Petit M, Grandbastien MA: **Distribution dynamics of the Tnt1 retrotransposon in tobacco.** *Mol Genet Genomics* 2007, **278(6)**:639-51.
48. Gao X, Hou Y, Ebina H, Levin HL, Voytas DF: **Chromodomains direct integration of retrotransposons to heterochromatin.** *Genome Res* 2008, **18**:359-369.
49. Benjak A, Forneck A, Casacuberta JM: **Genome-wide analysis of the "cut-and-paste" transposons of grapevine.** *PLoS One* 2008, **3(9)**:e3107.
50. **Phytozome: a tool for green plant comparative genomics.** [http://www.phytozome.net/].
51. Friedman AR, Baker BJ: **The evolution of resistance genes in multi-protein plant resistance systems.** *Curr Op Genet Dev* 2007, **17**:493-499.
52. Hernández-Pinzón I, de Jesús E, Santiago N, Casacuberta JM: **The frequent transcripcional readthrough of the tobacco Tnt1 retrotransposon and its possible implications for the control of resistance genes.** *J Mol Evol* 2009, **68(3)**:269-78.
53. Estill JC, Bennetzen JL: **The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes.** *Plant Methods* 2009, **5**:8.
54. McCarthy EM, McDonald JF: **LTR_STRUC: a novel search and identification program for LTR retrotransposons.** *Bioinformatics* 2003, **19(3)**:362-367.
55. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons.** *Nucleic Acids Res* 2007, **35**:W265-268.
56. Kalyanaraman A, Aluru S: **Efficient algorithms and software for detection of full-length retrotransposons.** *J Bioinform Comput Biol* 2006, **4(2)**:197-216.
57. **Repbase.** [http://www.girinst.org/repbase/index.html].
58. **National Center for Biotechnology Information.** [http://www.ncbi.nlm.nih.gov/].
59. Lewis SE, Searle SMJ, Harris H, Gibson M, Iyer V, Richter J, Wiel C, BAyraktaroglu L, Birney E, Crosby MA, Kaminker JS, Matthews B, Prochnik SE, Smith CD, Tupyl JL, Rubin GM, Misra S, Mungall CJ, Clamp ME: **Apollo: a sequence annotation editor.** *Genome Biology* 2002, **3(12)**:RESEARCH0082.
60. Cannon SB, Sterck L, Rombauts S, Sato S, cheung F, Gouzy G, Wang X, Mudge J, Vasdewani J, Scheix T, Spannagl M, Monaghan E, Nicholson C, Humphray SJ, Schoof H, Mayer KFX, Rogers J, Quetier F, Oldroyd GE, Debelle F, Cook DR, Retzel EF, Roe BA, Town CD, Tabata S, Van de Peer Y, Young ND: **Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes.** *Proc Natl Acad Sci USA* 2006, **103**:14959-14964.

doi:10.1186/1471-2229-10-246
Cite this article as: González *et al.*: Sequencing of 6.7 Mb of the melon genome using a BAC pooling strategy. *BMC Plant Biology* 2010 **10**:246.

The genome of melon (*Cucumis melo* L.)

Jordi Garcia-Mas[a,1], Andrej Benjak[a], Walter Sanseverino[a], Michael Bourgeois[a], Gisela Mir[a], Víctor M. González[b], Elizabeth Hénaff[b], Francisco Câmara[c], Luca Cozzuto[c], Ernesto Lowy[c], Tyler Alioto[d], Salvador Capella-Gutiérrez[c], Jose Blanca[e], Joaquín Cañizares[e], Pello Ziarsolo[e], Daniel Gonzalez-Ibeas[f], Luis Rodríguez-Moreno[f], Marcus Droege[g], Lei Du[h], Miguel Alvarez-Tejado[i], Belen Lorente-Galdos[j], Marta Melé[c,j], Luming Yang[k], Yiqun Weng[k,l], Arcadi Navarro[j,m], Tomas Marques-Bonet[j,m], Miguel A. Aranda[f], Fernando Nuez[e], Belén Picó[e], Toni Gabaldón[c], Guglielmo Roma[c], Roderic Guigó[c], Josep M. Casacuberta[b], Pere Arús[a], and Pere Puigdomènech[b,1]

[a]Institut de Recerca i Tecnologia Agroalimentàries, Centre for Research in Agricultural Genomics Consejo Superior de Investigaciones Científicas-Institut de Recerca i Tecnologia Agroalimentàries-Universitat Autònoma de Barcelona-Universitat de Barcelona, 08193 Barcelona, Spain; [b]Centre for Research in Agricultural Genomics Consejo Superior de Investigaciones Científicas-Institut de Recerca i Tecnologia Agroalimentàries-Universitat Autònoma de Barcelona-Universitat de Barcelona, 08193 Barcelona, Spain; [c]Centre for Genomic Regulation, Universitat Pompeu Fabra, 08003 Barcelona, Spain; [d]Centre Nacional d'Anàlisi Genòmica, 08028 Barcelona, Spain; [e]Institute for the Conservation and Breeding of Agricultural Biodiversity, Universitat Politècnica de Valencia, 46022 Valencia, Spain; [f]Departamento de Biología del Estrés y Patología Vegetal, Centro de Edafología y Biología Aplicada del Segura, Consejo Superior de Investigaciones Científicas, 30100 Murcia, Spain; [g]Roche Diagnostics Deutschland GmbH, 11668305 Mannheim, Germany; [h]Roche Diagnostics Asia Pacific Pte. Ltd., Singapore 168730; [i]Roche Applied Science, 08174 Barcelona, Spain; [j]Institut de Biologia Evolutiva, Universitat Pompeu Fabra-Consejo Superior de Investigaciones Científicas, 08003 Barcelona, Spain; [k]Horticulture Department, University of Wisconsin, Madison, WI 53706; [l]US Department of Agriculture–Agricultural Research Service, Horticulture Department, University of Wisconsin, Madison, WI 53706; and [m]Institució Catalana de Recerca i Estudis Avançats, 08010 Barcelona, Spain

Edited* by David C. Baulcombe, University of Cambridge, Cambridge, United Kingdom, and approved June 8, 2012 (received for review April 2, 2012)

We report the genome sequence of melon, an important horticultural crop worldwide. We assembled 375 Mb of the double-haploid line DHL92, representing 83.3% of the estimated melon genome. We predicted 27,427 protein-coding genes, which we analyzed by reconstructing 22,218 phylogenetic trees, allowing mapping of the orthology and paralogy relationships of sequenced plant genomes. We observed the absence of recent whole-genome duplications in the melon lineage since the ancient eudicot triplication, and our data suggest that transposon amplification may in part explain the increased size of the melon genome compared with the close relative cucumber. A low number of nucleotide-binding site–leucine-rich repeat disease resistance genes were annotated, suggesting the existence of specific defense mechanisms in this species. The DHL92 genome was compared with that of its parental lines allowing the quantification of sequence variability in the species. The use of the genome sequence in future investigations will facilitate the understanding of evolution of cucurbits and the improvement of breeding strategies.

de novo genome sequence | phylome

Melon (*Cucumis melo* L.) is a eudicot diploid plant species ($2n = 2x = 24$) of interest for its specific biological properties and for its economic importance. It belongs to the Cucurbitaceae family, which also includes cucumber (*Cucumis sativus* L.), watermelon [*Citrullus lanatus* (Thunb.) Matsum. & Nakai], and squash (*Cucurbita* spp.). Although originally thought to originate in Africa, recent data suggest that melon and cucumber may be of Asian origin (1). With its rich variability in observable phenotypic characters, melon was the inspiration for theories which were the precursors of modern genetics (2). Melon is an attractive model for studying valuable biological characters, such as fruit ripening (3), sex determination (4, 5), and phloem physiology (6).

Melon is an important fruit crop, with 26 million tons of melons produced worldwide in 2009 (http://faostat.fao.org). It is particularly important in Mediterranean and East Asian countries, where hybrid varieties have a significant and growing economic value. In line with the scientific and economic interest of the species, a number of genetic and molecular tools have been developed over the last years, including genetic maps (7), ESTs (http://www.icugi.org), microarrays (8), a physical map (9), BAC sequences (10), and reverse genetic tools (11, 12). To complete the repertoire of genomic tools, de novo sequencing of the melon genome was undertaken with 454 pyrosequencing. The genome sequence enabled an exhaustive phylogenic comparison of the melon genome with cucumber and other plant species. The melon and cucumber genome sequences are excellent tools for understanding the genome structure and evolution of two important species of the same genus with different chromosome number (melon, $2n = 2x = 24$; cucumber, $2n = 2x = 14$).

Results

Sequencing and Assembly of the Genome. The homozygous DHL92 double-haploid line, derived from the cross between PI 161375 (Songwhan Charmi, spp. *agrestis*) (SC) and the "Piel de Sapo" T111 line (ssp. *inodorus*) (PS), was chosen to obtain a better assembly of the genome sequence. A whole-genome shotgun strategy based on 454 pyrosequencing was used, producing 14.8 million single-shotgun and 7.7 million paired-end reads. Additionally, 53,203 BAC end sequences were available (13). After filtering the mitochondrial and chloroplast genomes (14), 13.52× coverage of the estimated 450-Mb melon genome (15) was obtained (*SI Appendix*, Table S1). Both 454 and Sanger reads were assembled with Newbler 2.5 into 1,594 scaffolds and 29,865 contigs, totaling 375 Mb of assembled genome (Table 1; *SI Appendix*, SI Text). The N50 scaffold size was 4.68 Mb, and 90% of the assembly was contained in 78 scaffolds (*SI Appendix*, Table S2). The assembly was corrected in homopolymer regions with Illumina reads. The melon genome assembly can be considered of good quality compared with other sequenced plant genomes based on next-generation sequencing (NGS) (*SI Appendix*, Table S3). We identified a considerable fraction (90.4%) of the unassembled reads as repeats containing transposable elements and low-complexity sequences. The difference between the estimated and the assembled genome size could be due to unassembled regions of repetitive DNA, similar to what has been found in genomes obtained with NGS (16).

Author contributions: J.G.-M., M.A.A., F.N., B.P., T.G., G.R., R.G., J.M.C., P.A., and P.P. designed research; A.B., W.S., M.B., G.M., V.M.G., E.H., F.C., L.C., E.L., T.A., S.C.-G., J.C., P.Z., D.G.-I., L.R.-M., M.D., L.D., M.A.-T., B.L.-G., M.M., L.Y., and Y.W. performed research; W.S., V.M.G., E.H., F.C., L.C., E.L., T.A., S.C.-G., J.B., A.N., T.M.-B., M.A.A., B.P., T.G., G.R., and J.M.C. analyzed data; and J.G.-M. and P.P. wrote the paper.

Conflict of interest statement: L.D., M.D., and M.A.-T. are Roche employees, and the work was partly funded by Roche.

Data deposition: The sequence data from this study have been deposited in the ENA Short Read Archive under accession no. ERP001463 and in the EMBL-Bank project PRJEB68. Further information is accessible through the MELONOMICS website (http://melonomics.net).

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence may be addressed. E-mail: jordi.garcia@irta.cat or pere.puigdomenech@cragenomica.es.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1205415109/-/DCSupplemental.

Table 1. Metrics of the melon genome assembly

Assembly	Measure
Bases in contigs	335,385,220
No. of contigs (>100 bases)	60,752
No. of large contigs (>500 bases)	40,102
Average large contig size (bases)	8,233
N50 large contig size (bases)	18,163
No. of scaffolds	1,594
Bases in scaffolds (including gaps)	361,410,028
No. of contigs in scaffolds	30,887
No. of bases in contigs in scaffolds	321,933,769
Average scaffold size (bases)	226,731
N50 scaffold size (bases)	4,677,790

The quality of the assembly was assessed by mapping it to four BACs that were previously sequenced using a shotgun Sanger approach. Overall, 92.5% of the BAC sequences were well represented in the genome assembly, aligning contiguously and with more than 99% similarity (*SI Appendix*, Fig. S1 and Table S4). The main source of error corresponded to gaps in the assembly located where transposons were annotated in the BAC sequences (*SI Appendix*, Table S5). A set of 57 BACs sequenced with 454 using a pooling strategy (10) was also compared with the assembly, which confirmed 92.3% of the BAC assemblies as being consistent with the genome assembly (*SI Appendix*, Table S6). The coverage of the melon genome was assessed by mapping 112,219 melon unigenes (17), of which 95.6% mapped unambiguously in the assembly, confirming a high level of coverage of the gene space.

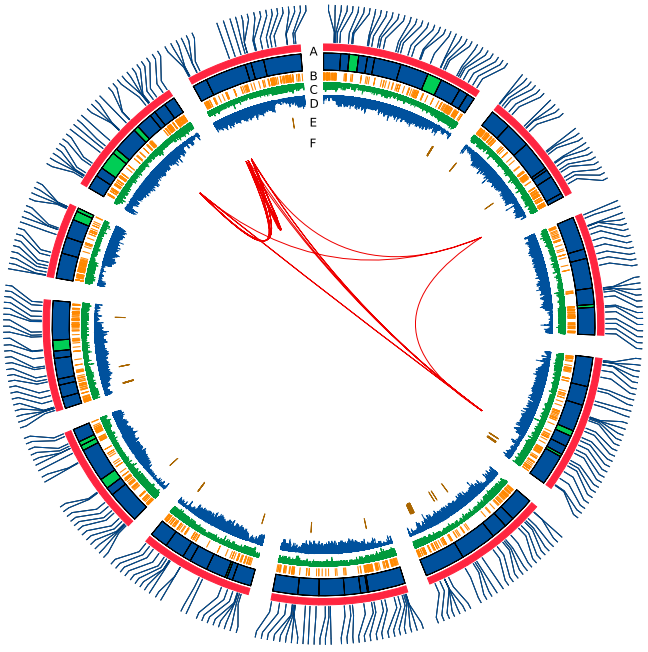


Fig. 1. The DHL92 melon genome. (*A*) Physical map of the 12 melon pseudochromosomes, represented clockwise starting from center above. Blocks represent scaffolds anchored to the genetic map. Scaffolds without orientation are in green. The physical location of SNP markers from the SC × PS genetic map is represented. (*B*) Distribution of ncRNAs (orange). (*C*) Distribution of predicted genes (light green). (*D*) Distribution of transposable elements (blue). (*E*) Distribution of NBS–LRR R-genes (brown). (*F*) Melon genome duplications. Duplicated blocks are represented as dark-green connecting lines.

Anchoring the Genome to Pseudochromosomes. A genetic map based on the SC × PS doubled haploid line mapping population, containing 602 SNPs, was used to anchor the assembly to 12 pseudochromosomes (*SI Appendix, Fig. S2*). We anchored 316.3 Mb of sequence contained in 87 scaffolds, representing 87.5% of the scaffold assembly (Fig. 1*A*; *SI Appendix*, Table S7). By anchoring the genetic map, we detected five scaffolds that mapped in two genomic locations due to misassemblies, which were manually corrected. The ratio between genetic and physical distances localized a region of recombination suppression in each pseudochromosome, which may correspond to the position of the centromeres (*SI Appendix*, Fig. S3).

Transposon Annotation. By using homology and structure-based searches, we identified 323 transposable element representatives belonging to the major superfamilies previously described in plants. These were used as queries to annotate 73,787 copies in the assembly, totaling 19.7% of the genome space. This percentage is similar to the one reported for genomes of similar size such as cacao (18). However, it is probably an underestimate as a result of the high stringency of our searches and the presence of additional transposon sequences in the unassembled fraction of the genome. The retrotransposon elements account for 14.7% of the genome whereas DNA transposons represent an additional 5.0% (*SI Appendix*, Table S8). A total of 87% of the annotated transposon-related sequences were attributed to a particular superfamily of elements and further classified into families. The transposable elements showed a complementary distribution to the gene space, probably representing the heterochromatic fraction (Fig. 1 *C* and *D*).

The two LTRs of LTR retrotransposons are identical upon insertion, and the number of differences between them can be used to determine the age of the insertion. We dated the insertion time of all LTR retrotransposons belonging to families containing at least 10 complete elements by intraelement comparison of LTRs (*SI Appendix*, SI Text). This analysis showed that, although different families had distinct patterns of amplification over time, most retrotransposons were inserted recently, with a peak of activity around 2 million years ago (Mya) (Fig. 2; *SI Appendix*, Fig. S4). As melon and cucumber ancestors diverged 10.1 Mya (1), our results suggest that high retrotransposition activity occurred in the melon lineage after this divergence. We applied the same annotation pipeline to look for retrotransposons in the Gy14 cucumber genome (http://www.phytozome.net) and found elements accounting for 1.5% of the genome. When less-stringent parameters were used, the percentage reached 4.8%, which was still significantly lower than the genome fraction annotated in melon, suggesting that LTR-retrotransposon activity was much higher and more recent in the melon lineage. Similar results were obtained when the annotation pipeline was applied to the 9930 cucumber genome (19). To assess whether DNA transposons have also been more active in the melon lineage than that of cucumber, we annotated in the Gy14 cucumber genome the three most represented superfamilies in both species (i.e., CACTA, MULE, and PIF/Harbinger) (*SI Appendix*, Table S8) (19), showing that all three have been amplified in the melon lineage (10× for CACTA, 47× for MULE, and 3.8× for PIF) (*SI Appendix*, Table S9).

Gene Prediction and Functional Annotation. The annotation of the assembled genome after masking repetitive regions resulted in a prediction of 27,427 genes with 34,848 predicted transcripts encoding 32,487 predicted polypeptides (*SI Appendix*, Table S10). Genes were preferentially distributed near the telomeres for most of the chromosomes (Fig. 1*C*). The average gene size for melon is 2,776 bp, with 5.85 exons per gene, similar to *Arabidopsis* (20), and a density of 7.3 genes per 100 kb, similar to grape (21). A total of 16,120 genes (58.7%) had exons supported by ESTs, and 14,337 (52.2%) were supported by GeneWise protein alignments, totaling 18,948 genes (69.1%) supported by a transcript and/or a protein alignment. The predicted melon proteins were annotated using an automatic pipeline. For each

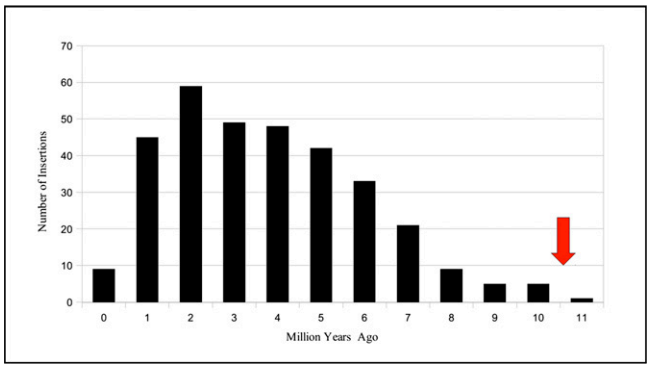


Fig. 2. LTR retrotransposon insertion during melon genome evolution. All LTR retrotransposon families with 10 or more copies were considered. Combined number of insertions for all families is displayed. Red arrow indicates when the melon and cucumber lineages diverged.

protein sequence, our approach identified protein signatures (*SI Appendix*, Table S11), assigned orthology groups, and used orthology-derived information to annotate metabolic pathways, multienzymatic complexes, and reactions.

Phylogenomic Analysis of Melon Across Other Plant Species. To assess the evolutionary relationships of melon genes in relation to other sequenced plant genomes, we undertook a comprehensive phylogenomic approach, which included reconstruction of the complete collection of evolutionary histories of all melon protein-coding genes across a phylogeny of 23 sequenced plants (i.e., the phylome; *SI Appendix*, Table S12). The usefulness of this approach in the annotation of newly sequenced genomes has been demonstrated in other eukaryotes (22, 23). A total of 22,218 maximum-likelihood (ML) phylogenetic trees were reconstructed and deposited at PhylomeDB (24) (http://phylomedb.org). We scanned the melon phylome to derive a complete catalog of phylogeny-based orthology and paralogy relationships across plant genomes (25). In addition, we used a topology-based approach (26) to detect and date duplication events. The alignments of 60 gene families with one-to-one orthology relationships across most plants were concatenated into a single alignment and used to derive a ML tree representing the evolutionary relationships of the species considered. The resulting topology was fully congruent with that obtained with the entire melon phylome using a gene tree parsimony approach, which minimizes the total number of inferred duplication events (27) (Fig. 3). Our phylogenetic analysis is in agreement with the assignment of *Populus* in the Malvidae clade (28).

Duplication analysis on entire phylomes has been used to confirm ancient whole-genome duplication (WGD) events, which emerge as duplication peaks in the corresponding evolutionary periods (29). Our results are consistent with the absence of WGD in the lineages leading to *C. melo*. Nevertheless, our approach detects several gene families that expanded specifically in the *Cucumis* and *C. melo* lineages. Duplicated genes are enriched in some functional processes, such as alcohol metabolism and defense response in the *Cucumis* lineage or phytochelatin metabolism and defense response in *C. melo* (Dataset S1). Expanded genes in the defense response and apoptosis functional processes belong to the coiled-coil (CC)–nucleotide-binding site (NBS)–leucine-rich repeat (LRR) (CNL) and toll/interleukin-1 receptor (TIR)-NBS-LRR (TNL) classes of disease resistance genes. The genes expanded in the phytochelatin metabolism functional process encode for phytochelatin synthase, an enzyme involved in resistance to metal poisoning. The genes expanded in the alcohol metabolism functional process encode (*R*)-(+)-mandelonitrile lyase, an enzyme involved in cyanogenesis, a defense system against herbivores and bacteria, the activity of which has been reported in melon seed (30). These expansions provide useful clues to establishing genetic links to the phenotypic particularities of these species.

Annotation of RNA Genes. A total of 1,253 noncoding RNA (ncRNA) genes were identified in the melon genome, similar to *Arabidopsis* (*SI Appendix*, Table S13; Dataset S2). In contrast to *Arabidopsis*, the ncRNA genes were distributed in the gene space (Fig. 1*B*). A total of 102 ncRNA were identified as forming 26 potential clusters (*SI Appendix*, Table S14). Of the 140 potential *MIRNA* loci identified, 122 corresponded to 35 known plant microRNA (miRNA) families, and expression data of mature miRNA sequences existed for at least 87 of them (31). Predicted precursors had an average size of 156 nt, ranging from 90 to 583 nt (Dataset S3). From a total of 19 *MIR169* members identified, 12

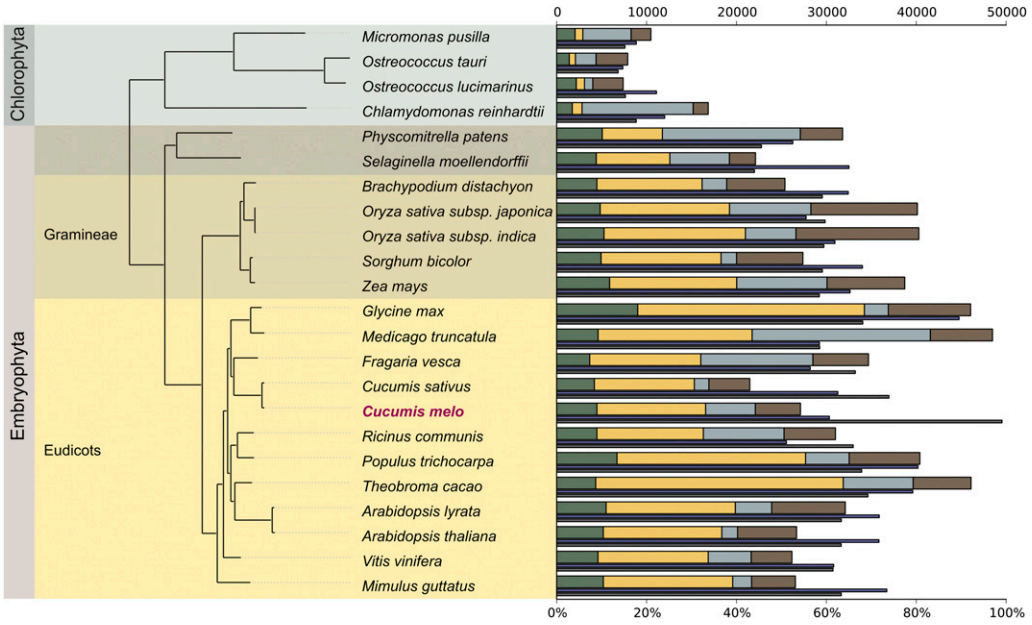


Fig. 3. Comparative genomics of 23 fully sequenced plant species where phylogeny is based on maximum-likelihood analysis of a concatenated alignment of 60 widespread single-copy proteins. Different background colors indicate taxonomic groupings within the species used to make the tree. Bars represent the total number of genes for each species (scale on the top). Bars are divided to indicate different types of homology relationships. Green: widespread genes that are found in at least 25 of the 28 species, including at least one out-group. Orange: widespread but plant-specific genes that are found in at least 20 of the 23 plant species. Gray: Species-specific genes with no (detectable) homologs in other species. Brown: genes without a clear pattern. The thin purple line under each bar represents the percentage of genes with a least one paralog in each species. The thin dark gray line represents the percentage of melon genes that have homologs in a given species.

were located in the same scaffold in a range of ~35 kb. Eight of them were found in pairs in a range of around 300 bases in the same DNA strand (*SI Appendix*, Fig. S5), suggesting simultaneous transcription in a single polycistronic transcript.

Disease Resistance Genes. A total of 411 putative disease resistance R-genes (32) were identified in the melon genome (*SI Appendix*, Table S15). Of these, 81 may exert their disease resistance function as cytoplasmatic proteins through canonical resistance domains, such as the NBS, the LRR, and the TIR domains (Fig. 1*E*). In addition, 290 genes were classified as transmembrane receptors, including 161 receptor-like kinases (RLK), 19 kinases containing an additional antifungal protein ginkbilobin-2 domain (RLK-GNK2), and 110 receptor-like proteins. Finally, 15 and 25 genes were found to be homologs to the barley *Mlo* (33) and the tomato *Pto* (34) genes, respectively. The number of R-genes in melon was found to be significantly lower than in other species. In cucumber and papaya, 61 and 55 genes from the cytoplasmic class were annotated, respectively, in contrast to 212 in *Arabidopsis* and 302 in grape. These data suggest that the number of NBS–LRR genes is not conserved among plant species and that the value is rather low in *Cucumis*, further suggesting a similar evolution of the NBS–LRR gene repertoire in these species.

R-genes were nonrandomly distributed in the melon genome, but organized in clusters (*SI Appendix*, Fig. S6; Dataset S4). In particular, 79 R-genes were located within 19 genomic clusters, 16 with genes belonging to the same family. This is a further indication that these genes are under rapid and specific evolution, with a strong tandem duplication activity. Overall, 45% of the NBS-LRR genes were grouped within nine clusters, whereas, in contrast, only 15% of the transmembrane receptors were clustered. Four clusters containing 13 TNL genes and spanning a region of 570 kb are located in the same region of the melon *Vat* resistance gene (35). Another cluster with seven TNL genes spanning 135 kb colocalized with the region harboring the *Fom-1* resistance gene (36). A cluster of six CNL genes spanning 56 kb and not described previously was located in LG I. The reconstructed phylogenies of some of these families revealed interesting scenarios: three lineage-specific independent RLK expansions involving several rounds of tandem duplications at three corresponding ancestral loci were identified (*SI Appendix*, Fig. S7). All members of each phylogenetic clade are located in the same genomic interval of less than 20 kb: two RLK genes in scaffold0008, three in scaffold0011, and four in scaffold0014. The same type of gene expansion was found for TNL genes from the cluster in scaffold00051 in LG IX, suggesting that there was amplification of an ancestral gene leading to the current cluster of R-genes in this genomic interval.

Genes Involved in Fruit Quality. Taste, flavor, and aroma of different melon types are the consequence of the balanced accumulation of many compounds. Among the major processes that occur during fruit ripening, two are particularly interesting from the breeding point of view: accumulation of sugars, which is responsible for the characteristic sweet taste, and carotenoid accumulation, which is responsible for the flesh color. Sixty-three genes putatively involved in sugar metabolism were annotated, belonging to 16 phylogenetic groups (Dataset S5). Twenty-one of these genes were not previously reported in melon (37, 38), of which 8 had EST support. A gene putatively encoding a UDP-glc phyrophosphorylase (*CmUGP-LIKE1*), for which a single gene was described (*CmUGP*), was annotated (*SI Appendix*, Fig. S8). A cell-wall invertase (*CmCIN-LIKE1*) was annotated, probably resulting from the duplication of *CmCIN2* in the ancestor of melon and cucumber (*SI Appendix*, Fig. S9). *CmSPS-LIKE1* may correspond to a member of the third subgroup of sucrose-P synthases not yet reported in melon, which are closely related to *Arabidopsis AtSPS4F*. Twenty-six genes encoding 14 enzymes involved in the plant carotenoid pathway were annotated, corresponding to 11 phylogenetic groups (Dataset S6), and 20 of the genes were supported by ESTs. These genes will permit us to obtain insight into the mechanisms controlling sucrose and carotene accumulation in melon fruit flesh.

Genome Duplications. Analysis of the genome sequence of several plant genomes has highlighted the existence of two ancestral WGDs (39) before the diversification of seed plants and angiosperms. An additional paleo-hexaploidization event (γ) followed by lineage-specific WGDs has shaped the structure of eudicot genomes (40). Using 4,258 melon paralogs, we identified 21 paralogous syntenic blocks within the melon genome, with no trace of a recent WGD (Fig. 1*F*; *SI Appendix*, Table S16).

Recent segmental duplications (SD) were searched for by combining two different methods. The whole-genome shotgun sequence detection (WSSD) method (41), based on detecting excess depth-of-coverage when mapping whole-genome sequence reads against the assembly, predicted 12.66 Mb of duplicated content (*SI Appendix*, Table S17). The whole-genome assembly comparison (WGAC) strategy (42), based on self-comparison of the whole genome using BLAST pairwise genome analysis, identified 4.37 Mb of duplicated sequence in the assembly. The resulting intersection between WSSD and WGAC is a good measure of the quality of duplicated content in a given assembly, detecting both artifact duplications and general collapse. We found an excess of possible collapses in the assembly (11.63 Mb) as a result of its construction based on short reads (43). The total of duplicated sequences identified by depth of coverage could still be an underestimate, given that the genome is highly fractionated. However, both types of analysis support limited segmental duplications in the melon genome.

Syntenic Relationships Between Melon and Other Plant Genomes. Comparison of melon and cucumber synteny suggested an ancestral fusion of five melon chromosome pairs in cucumber and several inter- and intrachromosome rearrangements (19, 44). We performed an alignment of both genomes, which showed the high level of synteny at higher resolution, and it allowed detecting shorter regions of rearrangements among chromosomes not previously observed (Fig. 4*A*; *SI Appendix*, Table S18). Our analysis suggests that melon LG I corresponds to cucumber chromosome 7, but with several inversions and an increase in the total chromosome size (35.8 vs. 19.2 Mb) (Fig. 4*C*). Melon LG IV and LG VI were fused into cucumber chromosome 3, but with several rearrangements and a reduction in total size in cucumber (30.4 and 29.8 Mb vs. 39.7 Mb) (Fig. 4*B*). The first distal 8.5 and 5 Mb of melon LG IV and cucumber chromosome 3, respectively, are highly collinear but with a progressive increase in size in melon toward the heterochromatic fraction (Fig. 4*D*), correlating with a higher density of transposable elements and a lower density of gene fraction (Fig. 1). There are other examples of more complex chromosomal rearrangements, but the total number of small inversions cannot be easily determined due to lack of orientation of some scaffolds in both species. Further refinement of the physical maps and sequencing of other *Cucumis* species may shed light on the genome structure of the ancestor of cucumber and melon.

A total of 19,377 one-to-one ortholog pairs were obtained between melon and cucumber, yielding 497 orthologous syntenic blocks when using stringent parameters (*SI Appendix*, Table S19 and Fig. S10) and showing a similar pattern to that obtained after the complete genome alignments. The melon genome was also compared with the genomes of *Arabidopsis*, soybean, and *Fragaria vesca*, on the basis of the orthologous genes identified in the phylome analysis. *Fragaria*, melon, and soybean belong to the Fabidae clade, whereas *Arabidopsis* is in the Malvidae clade. Two rounds of WGD have been reported for *Arabidopsis* and soybean, whereas no WGD has been found in *Fragaria*. We found a higher number of synteny blocks with soybean and *Fragaria* than with *Arabidopsis* (*SI Appendix*, Table S19 and Fig. S10).

DHL92 Genome Structure Based on Resequencing Its Parental Lines. DHL92 and its parental lines SC and PS were resequenced using

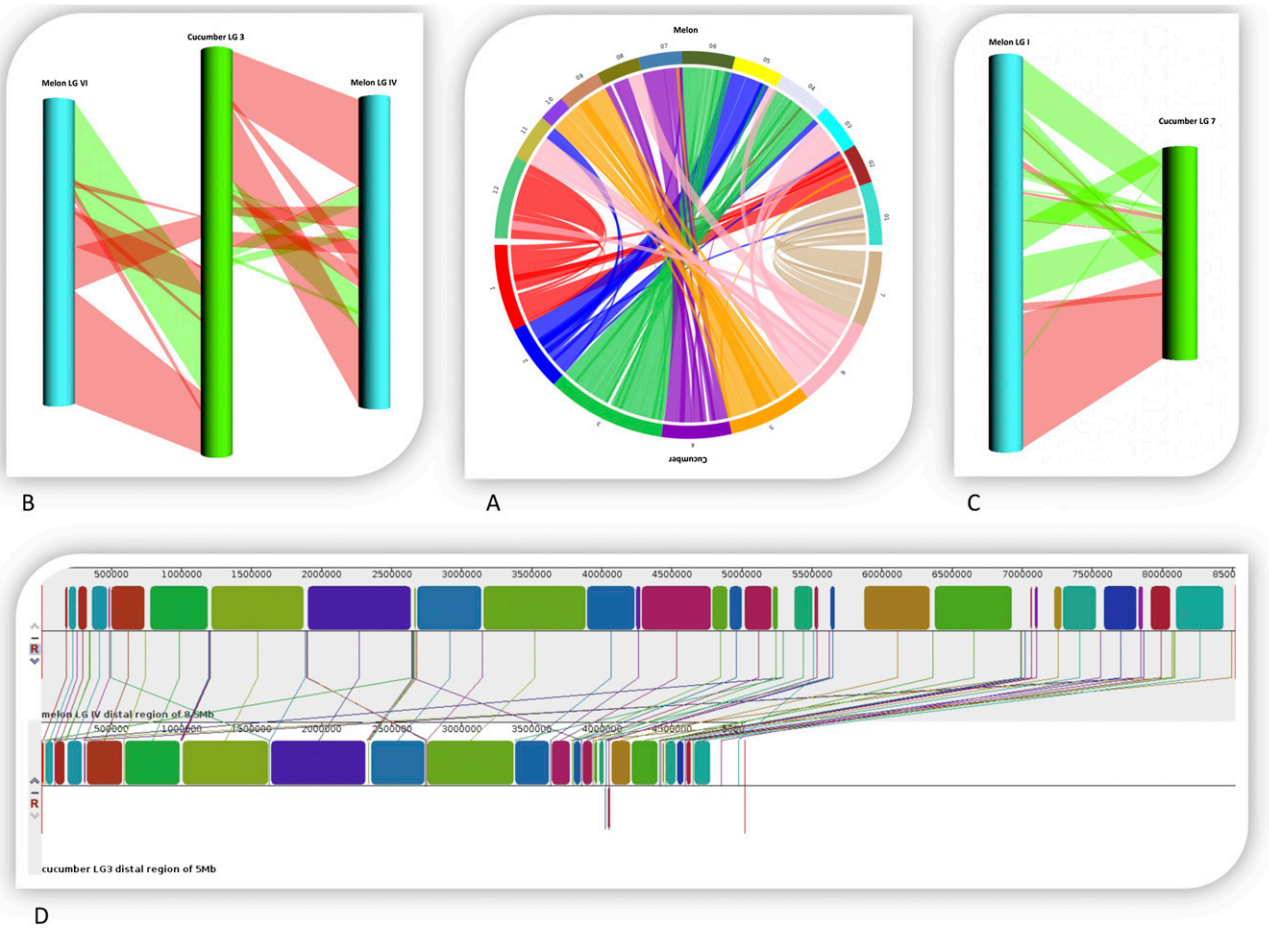


Fig. 4. Comparative analysis of the melon and cucumber genomes. (*A*) Alignment of melon (x = 12) and cucumber (x = 7) genomes. (*B*) Alignment of melon LG IV and LG VI with cucumber chromosome 3. Direct blocks are represented in red and inverted blocks in green. (*C*) Alignment of melon LG I with cucumber chromosome 7. Direct blocks are represented in red and inverted blocks in green. (*D*) Genome expansion in melon LG IV distal region of 8.5 Mb (*Upper*) compared with cucumber chromosome 3 distal region of 5 Mb (*Lower*). Blocks of the same color correspond to syntenic regions.

the Illumina GAIIx platform, yielding 213 million 152-bp reads (*SI Appendix*, Table S20), which were aligned to the DHL92 reference genome. We identified 2.1 million SNPs and 413,000 indels between DHL92 and both parental lines (*SI Appendix*, Table S21), from which 4.0% and 3.1% were located in exons, respectively. We could reconstruct the DHL92 genome on the basis of its parental lines (*SI Appendix*, Fig. S11 and S12), which contain a total of 17 recombination events, with an average of 1.4 recombinations per linkage group. The number of SNPs and indels between SC and PS resulted in a frequency of one SNP every 176 bp and one indel every 907 bp.

Discussion

The increasing availability of genome sequences from higher plants provides us with an important tool for understanding plant evolution and the genetic variability existing within cultivated species. Genome sequences are also becoming a strategic tool for the development of methods to accelerate plant breeding. The Cucurbitaceae is, after the Solanaceae, the most economically important group of vegetable crops, especially in Mediterranean countries. Melon has a key position in the Cucurbitaceae family for its high economic value and as a model to study biologically relevant characters, so the melon genome sequence has the added value of providing breeders with an additional tool in breeding programs. For these reasons, the availability of a good-quality draft sequence of the melon genome is essential.

The combination of different sequencing strategies and the use of a double-haploid line were important factors for assembling the genome in large scaffolds (N50 scaffold size 4.68 Mb). This gave a high-quality genome assembly compared with some of the recently published plant genomes that used NGS technologies. The quality of the assembly has an impact on further uses of the genome sequence, providing an efficient reference genome for resequencing analysis. The resequencing of the parents of the DHL92 reference genome allowed a first measure of the polymorphism in melon, as more than 2 million putative SNPs were identified.

The annotation of the assembled genome predicted 27,427 genes, a number similar to other plant species. A phylogenetic analysis of gene families greatly helped in the quality of the prediction. The number of predicted R-genes in melon and cucumber was lower than in other plant species. Expansion of the lipoxygenase gene family has been suggested as a complementary mechanism to challenge biotic stress in cucumber (19), but we did not observe such an expansion in melon. Therefore, the low number of R-genes in Cucurbitaceae may be the consequence of a different adaptive strategy of these species, which may be related to specific mechanisms of regulation of disease resistance genes or to their characteristic vascular structure (6). The availability of the genome sequence will be very valuable in studying this question that is also of importance for breeding biotic resistance.

Increase in genome size may, in general, be attributed to transposable element amplification and to polyploidization. Our analysis suggests that the melon genome did not have any recent lineage-specific whole-genome duplication, as in cucumber (19).

The closest families to cucurbits in the Fabidae clade are the Rosaceae, which includes species such as apple where a recent WGD has occurred; strawberry with no observable WGD; and Fabaceae, which includes species that share a recent WGD (soybean, *Medicago*, *Lotus*). As the number of available plant genomes increases, the observation of WGD events will help to understand their evolution. In cucurbits, the genome sequence of additional species will determine whether the lack of a recent WGD is unique to this lineage. Traces of duplications observed in melon may correspond to the ancestral paleo-hexaploidization that occurred after the divergence of monocots and dicots (40), with subsequent genome rearrangements and genome size reduction. Transposable elements have accumulated to a greater extent in melon compared with cucumber with a peak of activity around 2 Mya, suggesting that the larger genome size of melon, probably to a large extent, may be due to transposon amplification. However, loss of chromosome fragments during chromosome fusion in cucumber may also explain the larger melon genome. Melon and cucumber diverged only around 10 million years ago and are interesting models for studying genome size and chromosome number evolution (450 vs. 367 Mb and $x = 12$ vs. $x = 7$). We have shown that our sequence may be a good reference for resequencing other melon varieties. Further resequencing of other melon lines representing the extant variability of the species will also permit identification of SNPs and indels that may be used in breeding programs and in studying the genome rearrangements that have shaped the present structure of cucurbit genomes.

Materials and Methods

The melon doubled-haploid line DHL92 was derived from the cross between the Korean accession PI 161375 (Songwhan Charmi, spp. *agrestis*) (SC) and the "Piel de Sapo" T111 line (ssp. *inodorus*) (PS). DHL92 was chosen for its homozygosity. See *SI Appendix* for details of sequencing, assembly, annotation, and genome analysis.

ACKNOWLEDGMENTS. We thank Marc Oliver (Syngenta) for the recombinant inbred line genetic map. The cucumber Gy14 genome was produced by the Joint Genome Institute (http://www.jgi.doe.gov/). We acknowledge funding from Fundación Genoma España; Semillas Fitó; Syngenta Seeds; the governments of Catalunya, Andalucía, Madrid, Castilla-La Mancha, and Murcia; Savia Biotech; Roche Diagnostics; and Sistemas Genómicos. P.P. and J.G.-M. were funded by the Spanish Ministry of Science and Innovation (CSD2007-00036) and the Xarxa de Referència d'R+D+I en Biotecnologia (Generalitat de Catalunya). R.G. and A.N. acknowledge the Spanish National Bioinformatics Institute for funding. T.M.-B. is supported by European Research Council Starting Grant StG_20091118.

1. Sebastian P, Schaefer H, Telford IRH, Renner SS (2010) Cucumber (Cucumis sativus) and melon (C. melo) have numerous wild relatives in Asia and Australia, and the sister species of melon is from Australia. *Proc Natl Acad Sci USA* 107:14269–14273.
2. Sageret A (1826) Considérations sur la production des hybrides, des variantes et des variétés en général, et sur celles de la famille des Cucurbitacées en particulier [Considerations on the production of hybrids, variants and varieties in general and those of the Cucurbitaceae family in particular]. *Annales des Sciences Naturelles* 8:294–314.
3. Pech JC, Bouzayen M, Latché A (2008) Climacteric fruit ripening: Ethylene-dependent and independent regulation of ripening pathways in melon fruit. *Plant Sci* 175: 114–120.
4. Boualem A, et al. (2008) A conserved mutation in an ethylene biosynthesis enzyme leads to andromonoecy in melons. *Science* 321:836–838.
5. Martin A, et al. (2009) A transposon-induced epigenetic change leads to sex determination in melon. *Nature* 461:1135–1138.
6. Zhang B, Tolstikov V, Turnbull C, Hicks LM, Fiehn O (2010) Divergent metabolome and proteome suggest functional independence of dual phloem transport systems in cucurbits. *Proc Natl Acad Sci USA* 107:13532–13537.
7. Díaz A, et al. (2011) A consensus linkage map for molecular markers and quantitative trait loci associated with economically important traits in melon (*Cucumis melo* L.). *BMC Plant Biol* 11:111.
8. Mascarell-Creus A, et al. (2009) An oligo-based microarray offers novel transcriptomic approaches for the analysis of pathogen resistance and fruit quality traits in melon (Cucumis melo L.). *BMC Genomics* 10:467.
9. González VM, Garcia-Mas J, Arús P, Puigdomènech P (2010) Generation of a BAC-based physical map of the melon genome. *BMC Genomics* 11:339.
10. González VM, et al. (2010) Sequencing of 6.7 Mb of the melon genome using a BAC pooling strategy. *BMC Plant Biol* 10:246.
11. Dahmani-Mardas F, et al. (2010) Engineering melon plants with improved fruit shelf life using the TILLING approach. *PLoS ONE* 5:e15776.
12. González M, et al. (2011) Towards a TILLING platform for functional genomics in Piel de Sapo melons. *BMC Res Notes* 4:289.
13. González VM, et al. (2010) Genome-wide BAC-end sequencing of *Cucumis melo* using two BAC libraries. *BMC Genomics* 11:618.
14. Rodríguez-Moreno L, et al. (2011) Determination of the melon chloroplast and mitochondrial genome sequences reveals that the largest reported mitochondrial genome in plants contains a significant amount of DNA having a nuclear origin. *BMC Genomics* 12:424.
15. Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. *Plant Mol Biol Rep* 9:208–218.
16. Xu X, et al.; Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature* 475:189–195.
17. Blanca J, et al. (2011) Melon transcriptome characterization. SSRs and SNPs discovery for high throughput genotyping across the species. *Plant Genome* 4:118–131.
18. Argout X, et al. (2011) The genome of *Theobroma cacao*. *Nat Genet* 43:101–108.
19. Huang S, et al. (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat Genet* 41: 1275–1281.
20. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.
21. Jaillon O, et al.; French-Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467.
22. International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol* 8:e1000313.
23. Huerta-Cepas J, Marcet-Houben M, Pignatelli M, Moya A, Gabaldón T (2010) The pea aphid phylome: A complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for *Acyrthosiphon pisum* genes. *Insect Mol Biol* 19 (Suppl 2):13–21.
24. Huerta-Cepas J, et al. (2011) PhylomeDB v3.0: An expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res* 39(Database issue):D556–D560.
25. Gabaldón T (2008) Large-scale assignment of orthology: Back to phylogenetics? *Genome Biol* 9:235.
26. Huerta-Cepas J, Gabaldón T (2011) Assigning duplication events to relative temporal scales in genome-wide studies. *Bioinformatics* 27:38–45.
27. Wehe A, Bansal MS, Burleigh JG, Eulenstein O (2008) DupTree: A program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* 24:1540–1541.
28. Shulaev V, et al. (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nat Genet* 43:109–116.
29. Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T (2007) The human phylome. *Genome Biol* 8:R109.
30. Hernández L, Luna H, Ruíz-Terán F, Vázquez A (2004) Screening for hydroxynitrile lyase activity in crude preparations of some edible plants. *J Mol Catal B-Enzym* 30: 105–108.
31. Gonzalez-Ibeas D, et al. (2011) Analysis of the melon (*Cucumis melo*) small RNAome by high-throughput pyrosequencing. *BMC Genomics* 12:393.
32. Sanseverino W, et al. (2010) PRGdb: a bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Res* 38(Database issue):D814–D821.
33. Büschges R, et al. (1997) The barley *Mlo* gene: A novel control element of plant pathogen resistance. *Cell* 88:695–705.
34. Loh Y-T, Martin GB (1995) The disease-resistance gene *Pto* and the fenthion-sensitivity gene *fen* encode closely related functional protein kinases. *Proc Natl Acad Sci USA* 92:4181–4184.
35. Lecoq H, Pitrat M (1982) Effect on cucumber mosaic virus incidence of the cultivation of partially resistant muskmelon cultivars. *Acta Hortic* 127:137–145.
36. Oumouloud A, Arnedo-Andres MS, Gonzalez-Torres R, Alvarez JM (2008) Development of molecular markers linked to the *Fom-1* locus for resistance to *Fusarium* race 2 in melon. *Euphytica* 164:347–356.
37. Dai N, et al. (2011) Metabolism of soluble sugars in developing melon fruit: A global transcriptional view of the metabolic transition to sucrose accumulation. *Plant Mol Biol* 76:1–18.
38. Clepet C, et al. (2011) Analysis of expressed sequence tags generated from full-length enriched cDNA libraries of melon. *BMC Genomics* 12:252.
39. Jiao Y, et al. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.
40. Paterson AH, Freeling M, Tang H, Wang X (2010) Insights from the comparison of plant genome sequences. *Annu Rev Plant Biol* 61:349–372.
41. Bailey JA, et al. (2002) Recent segmental duplications in the human genome. *Science* 297:1003–1007.
42. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE (2001) Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res* 11:1005–1017.
43. Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. *Nat Methods* 8:61–65.
44. Li D, et al. (2011) Syntenic relationships between cucumber (*Cucumis sativus* L.) and melon (*C. melo* L.) chromosomes as revealed by comparative genetic mapping. *BMC Genomics* 12:396.

Chapter 7
MITEs, Miniature Elements with a Major Role in Plant Genome Evolution

Hélène Guermonprez, Elizabeth Hénaff, Marta Cifuentes, and Josep M. Casacuberta

Abstract Miniature Inverted-repeat Transposable Elements (MITEs) are a particular type of class II transposons found in genomes in high copy numbers. Most MITEs are deletion derivatives of class II transposons whose transposases have been shown to mobilize them by a typical cut-and-paste mechanism. However, unlike class II transposons, MITEs can amplify rapidly and dramatically and attain very high copy numbers, in particular, in plant genomes. This high copy number, together with their close association with genes, endows MITEs with a high potential to generate variability, and impact gene and genome evolution.

Keywords MITE-Class II transposons • Transposition mechanism • Impact of transposition • Amplification

Abbreviations

MITE	Miniature inverted-repeat transposable element
TE	Transposable element
TIRs	Terminal inverted repeats
TSD	Target site duplication

H. Guermonprez • E. Hénaff • J.M. Casacuberta (✉)
Department of Molecular Genetics, Center for Research in Agricultural Genomics, CRAG (CSIC-IRTA-UAB-UB), Campus UAB, Bellaterra (Cerdanyola del Vallés), 08193 Barcelona, Spain
e-mail: josep.casacuberta@cragenomica.es

M. Cifuentes
Institut Jean-Pierre Bourgin, UMR1318 INRA-AgroParisTech INRA Centre de Versailles-Grignon, Bâtiment 7, Route de St-Cyr (RD10), 78026 Versailles Cedex, France

M.-A. Grandbastien and J.M. Casacuberta (eds.), *Plant Transposable Elements*, Topics in Current Genetics 24, DOI 10.1007/978-3-642-31842-9_7,

7.1 Introduction

The term Miniature Inverted-repeat Transposable Elements (MITEs) was coined to designate different families of short mobile elements featuring Terminal Inverted Repeats (TIRs) and found in plant genomes in high copy number (Wessler et al. 1995). The first two families described were *Tourist* and the *Stowaway* from maize (Bureau and Wessler 1992, 1994). Sequence homology searches revealed their high similarity to transposons of *Mariner* and *PIF* families, respectively, suggesting that they could be deletion derivatives of class II transposons (Feschotte and Mouches 2000; Zhang et al. 2001). Since then, MITEs related to all major families of class II transposons have been reported (Benjak et al. 2009; Kuang et al. 2009; Yang and Hall 2003b), and MITE families have been described in both prokaryote and eukaryote genomes (Dufresne et al. 2007; Filee et al. 2007; Han et al. 2010; Piriyapongsa and Jordan 2007; Surzycki and Belknap 2000), including virtually all plant genomes analyzed (Benjak et al. 2009; Bergero et al. 2008; Bureau et al. 1996; Cantu et al. 2010; Casacuberta et al. 1998; Grzebelus et al. 2009; Kuang et al. 2009; Lyons et al. 2008; Momose et al. 2010; Sarilar et al. 2011; Schwarz-Sommer et al. 2010; Yang and Hall 2003b). However, while most MITEs seem to be deletion derivatives of autonomous elements, which probably mobilize them, in some cases the situation is less clear. Some MITEs cannot be related to long coding elements suggesting that in some cases MITEs may arise by the serendipitous juxtaposition of two inverted repeated sequences which may be recognized by an existing transposase (Feschotte and Pritham 2007). In other cases, like that of *mPing* in rice, the related long coding element has been identified but is absent from the varieties where *mPing* is active, suggesting that the element that gave rise to the MITE has been lost and that other transposases may catalyze its mobilization (Jiang et al. 2003). The emerging picture is thus a complex relationship between MITEs and their distantly related autonomous elements (Feschotte et al. 2005).

In addition to their small size and the presence of TIRs, a number of other characteristics have been associated with MITEs. The sequence of the first MITEs described was shown to be A/T-rich and to have the potential to form highly stable secondary structures (Bureau and Wessler 1992), and these characteristics seem to be shared by a high proportion of the MITEs described to date. However, during these years no evidence has demonstrated any relevance of these characteristics for MITEs' amplification dynamics.

MITEs are frequently found within or close to genes (Casacuberta and Santiago 2003), although this preference probably varies among different families (Mao et al. 2000). This trend, combined with their high copy number, endows MITEs with a great potential to modify gene expression upon mobilization (Deragon et al. 2008). In this chapter we summarize recent advances in the identification of MITEs, their mechanism of transposition, and their impact on genes and genomes. We also point out open questions regarding these miniature but highly complex elements.

7.2 MITE Identification

Due to their small size and absence of coding capacity, MITE identification and annotation is particularly difficult. As is the case for most TE (Transposable Element) families described to date, the first MITEs discovered were elements inserted in genes, causing a detectable mutation and phenotype. However, the availability of whole genome sequences together with the development of appropriate bioinformatic tools has enabled the discovery of the high prevalence of these elements in eukaryotic genomes.

7.2.1 Discovery by Insertional Mutagenesis

The first MITE, dubbed *Tourist*, was discovered in maize by insertional mutation in the *waxy* gene (Bureau and Wessler 1992). Its analysis revealed the presence of TIRs in the insert, which, combined with the fact that it was found in many copies in the available gene sequences of the same line, and the presence of a flanking duplicated sequence, led to the hypothesis that this was actually a mobile repeated element. Since then, other cases of insertional mutagenesis have led to the discovery of a few other MITEs such as *mPing* that was found inserted into the gene for *rice ubiquitin-related modifier-1* (*Rurm1*) and whose excision resulted in the reversion of the "slender glume" phenotype (Nakazaki et al. 2003) and *dTstu1*, the source of a somaclonal variation inducing purple pigment synthesis in a usually red potato variety (Momose et al. 2010).

7.2.2 Discovery by Bioinformatic Methods

While MITEs as a new superfamily of transposable elements were stumbled upon by accident and studied using molecular biology techniques, the availability of genomic sequence data as well as sequence search tools has allowed the identification of MITE families by bioinformatic means. One category of methods is based on sequence similarity to a known MITE or autonomous class II transposon. The second is to identify MITE families de novo, exploiting their structural characteristics and the fact that they are found in large copy numbers.

Certain MITE families are shared among several species, as is the case for *Tourist* in cereals, and can be detected by sequence similarity to already defined MITEs. For example, elements similar to the consensus sequences of the MITEs first identified in maize and barley (Bureau and Wessler 1992) were found in rice and sorghum (Bureau and Wessler 1994).

Many MITEs arise as deletion derivatives of their autonomous counterparts, and thus display sequence similarity to class II transposons. Exploiting this sequence similarity, new MITEs can be discovered by searching with a full-length element as

a query. However, while some MITEs are homologous to their autonomous counterparts in their entire length, others only share the TIR sequences and the rest of the internal sequence is unrelated, requiring different computational approaches for either case.

In the first case, MITEs can be identified by genome-wide similarity searches using the full-length TE as query, as was done in *Vitis vinifera* to identify MITEs related to known elements in the CACTA, hAT, and PIF superfamilies (Benjak et al. 2009).

The second case is more difficult as TIRs are short (10–20 nucleotides), and these can give many spurious hits. Various softwares have been developed to implement this search. A first example is TRANSPO that takes a TIR sequence and searches for inverted matches within a certain window and can be paired with the SPAT software, which performs a hierarchical clustering of the results, thus defining families of putative elements (Santiago et al. 2002). A second example is the MAK toolkit (Yang and Hall 2003a) that provides a suite of programs to identify MITE copies, or a related autonomous element, given a MITE query. This software implements various modes with different goals. The *Member Retriever* mode is designed to retrieve other MITEs similar to the supplied MITE query. The *Anchor* mode aims to identify autonomous elements that are related to a given MITE query, and the *Associator* mode reports gene annotations nearest to the hits.

With the recent proliferation of whole-genome sequencing data and comparative analyses, it becomes tempting to mine this wealth of information for entirely new MITEs using computational methods. Two different approaches for de novo MITE identification have been used to date, one based on comparative analyses of closely related organisms and the other exploiting the elements' structural characteristics and the fact that they are found in very high copy number.

The first approach is not specific to MITEs, but has lead to the identification of new MITE families in solanacae related to *hAT*, *Mutator*, *Stowaway* and *Tourist* elements by inspecting syntenic regions of resistance gene clusters in tomato, potato, and tobacco (Kuang et al. 2009). This method of searching for Related Empty Sites also provides indirect evidence for their mobilization, as discussed below.

The second approach is based on the fact that MITEs present very clear structural characteristics—exact TIRs and TSDs (target site duplication) upon insertion. However, these structures are very short and similar ones can arise by chance, leading to many false positives when the search criteria are limited to two inverted repeats flanked by direct ones. Thus, the true challenge of in silico MITE identification is eliminating false positives. Various programs have been developed for MITE identification in genomic sequences, the latest being MITE-hunter (Han and Wessler 2010). This software is the most sophisticated in that it provides several methods of eliminating false positives, at various steps of the algorithm. Similarly to others [FINDMITE (Tu 2001); MUST (Chen et al. 2009)], the first step is to identify candidate MITEs based on TIRs and TSDs. In a subsequent step, candidates are discriminated based on copy number by pairwise comparison—elements that do not align with any other are eliminated as false positives. Then a consensus sequence is generated for each family and the definition of its borders verified by multiple sequence alignment with its copies taken with flanking regions. This last step relies

on the fact that within a certain family, the copies' terminal sequences (i.e., TIRs and TSDs) will be near identical and align well but the alignment will break down at the flanking regions as each element is inserted in a different genomic context.

The surge of available genomic sequence data is a wealth of information for studying transposons in general, and MITEs in particular. Whole genome sequences provide the possibility of mining for new elements, impossible until the advent of this data. Also, comparative analyses between genomes are a powerful tool for identification of new elements and following TE movement. Two major technological advances, besides the progress in sequencing technologies, permit this: the development of algorithms for accurate whole-genome alignments (Frith et al. 2010) and genome resequencing (Stratton 2008). Until now transposon discovery by comparative analysis has been limited to certain syntenic regions, but exploiting this type of data on a whole genome scale is a promising prospect. Resequencing of varieties or lines within a species has the advantage of providing highly comparable data of closely related organisms, giving a perspective of the variations of the transposon landscape at a small evolutionary scale. Recently, the resequencing of rice lines issued from cell culture led to the identification of 43 new insertions of 13 different TEs. Although the authors have not exploited this analysis to look for new elements, their approach could also be used for de novo identification. In conclusion, genomic data analysis has provided evidence for MITE mobility and enabled the discovery of new elements. Furthermore, we can expect that the level of detail and precision at which we can study mobile elements on the genomic scale will increase with progress in algorithms for sequence analysis and quantity of data available.

7.3 MITE Transposition Mechanisms

The analysis of *Tourist*, the first MITE family characterized (Bureau and Wessler 1992), allowed for a first description of the particular characteristics of MITEs. *Tourist* elements presented TIRs and subterminal repeated sequences, as well as TSDs flanking the elements, which make them similar to class II transposons. However, these elements were present at a higher copy number than typical class II elements, and their copies showed an unprecedented homogeneity in size and sequence. These characteristics, later shown to be shared by most MITEs, made it difficult at the time to classify them. Moreover, MITEs' transposition mechanism remained a mystery as no excision event had yet been observed (Wessler et al. 1995).

The first evidence of MITEs' capacity for excision came from the phylogenetic analysis of the *Stowaway* family in 30 *Triticae* species (Petersen and Seberg 2000) and was later confirmed by the analysis of a rice *slender glume* mutant, which carries an *mPing* MITE whose excision lead to the reversion of the mutant phenotype (Nakazaki et al. 2003). The confirmation of MITEs' potential for excision, together with the fact that some show high sequence similarity with class II transposons (Feschotte and Mouches 2000), strongly suggested that MITEs could be deletion derivatives of class II transposons, mobilized by transposases encoded by their related

autonomous elements (Casacuberta and Santiago 2003; Feschotte et al. 2002; Zhang et al. 2001). This hypothesis gained further support from studies showing that the transposases encoded by class II transposons specifically bind the TIRs and subterminal sequences of related MITEs (Feschotte et al. 2005; Loot et al. 2006). The mobilization of MITEs by class II transposases was finally demonstrated in three independent reports in animals, plants, and fungi, which showed conclusive evidence that transposases from a related element were able to mobilize MITEs in vivo (Dufresne et al. 2007; Miskey et al. 2007; Yang et al. 2007). This mobilization has also been observed in heterologous systems (Hancock et al. 2010, 2011; Yang et al. 2007), suggesting that, as is the case for typical class II elements, the minimal requirements for MITEs transposition are a transposase and its binding sequences within the element. However, although MITEs' transposition seems in some respects very similar to that of typical class II elements, it also presents particular features that make MITEs a very unique type of defective class II elements.

First of all, MITEs seem to be particularly promiscuous with respect to the transposase they can use for mobilization. Phylogenetic analyses of rice *Mariner*-like elements and their related *Stowaway* MITEs suggested that homology restricted to the TIRs and subterminal sequences may be sufficient for cross-mobilization (Feschotte et al. 2003). This was confirmed by in vitro protein/DNA interaction studies showing that rice *Stowaway* MITEs can interact with transposases encoded by a panoply of *Mariner*-like *Osmar* elements (Feschotte et al. 2005). This promiscuity may explain the transposition of the rice *Tourist*-like element *mPing*, which is a deletion derivative of a class II element *Ping*, in rice cultivars that are devoid of active *Ping* elements but contain potentially active elements of the distantly related transposon *Pong* (Jiang et al. 2003). Indeed, recent experiments have demonstrated that *mPing* can be mobilized in vivo by both *Ping* and *Pong*'s transposases (Hancock et al. 2010). Based on these observations a model of MITE dynamics has been proposed in which MITEs would be generated through a deletion in an autonomous transposon, then amplification would take place maybe long afterwards, catalyzed by the element's encoded transposase or that of a distantly related element, as the former may even have disappeared (Jiang et al. 2004).

Second, some reports suggest that MITEs may be mobilized more efficiently than typical class II transposons. It has been shown that some transposases bind with higher affinity to the MITE sequence than to the transposase-encoding element, either because the MITE contains additional transposase binding sites in the subterminal repeated regions (Loot et al. 2006) or because it lacks repressive sequences present in the original autonomous element (Yang et al. 2009). Both MITEs' promiscuity and their higher transposase binding affinity could account for an increased transposition efficiency with respect to typical class II transposons. However, this does not seem to explain the third and most striking particularity of MITEs: their high copy number. Indeed, although the transposition process may in some cases lead to a moderate increase in copy number (as is the case for typical class II transposons), it is hard to imagine that the very high copy numbers MITEs can attain in very short evolutionary timescales (see below) are the result of an increased number of normal cut-and-paste transposition events. Moreover, while

MITEs do excise, excision events seem to be rare, as most MITE insertions are relatively stable even to the point of being used as genetic markers (Feschotte et al. 2002), suggesting that excisions do not correlate with MITE amplification.

What it is known to date explains how MITEs transpose but not how they amplify to the elevated copy numbers they usually reach in genomes. MITE transposition and amplification may be two different and uncoupled processes (Casacuberta and Santiago 2003; Feschotte et al. 2002) with the standard cut-and-paste transposition generating a moderate or no increase in copy number and amplification occurring rarely. Alternatively, amplification may result from transposition in particular cell types or conditions with higher DNA replication with respect to cell division, such as endoreduplicating cells.

A structural particularity of most MITEs for which a function has not yet been determined is their capacity to form highly stable single strand secondary structures. While it does not seem to be required for MITE cut-and-paste transposition (Sinzelle et al. 2008), it could affect MITEs amplification. It is tempting to hypothesize that the formation of single-strand hairpin structures, with double stranded TIRs, could allow transposase binding and single-stranded excision. It is interesting to note that the bacterial transposons of the IS200/IS605 family move by the excision and reintegration of only one of the strands of the transposon leaving the complementary strand behind. This mechanism is catalyzed by a very particular type of transposase and linked to replication (Guynet et al. 2008; Ton-Hoang et al. 2010). This particular mode of transposition could easily explain an increase of transposon copies. In plants, where endoreduplication or re-replication processes are commonplace, such a mechanism could be particularly relevant.

Irrespective of the mechanism responsible for MITEs amplification, their high copy number suggests that these elements are particularly successful in avoiding genome control. Interestingly MITEs are present at a much higher copy number than the elements coding for the transposase, which mobilize them and from which they frequently derive from. As silencing is the most general and efficient mechanism to control transposons (Lisch 2009), the separation of the transposase encoding element, which can be maintained at a low copy number and thus will not attract silencing, from the transposing unit, the MITE, more difficult to control as it does not need to be transcribed, could in part explain their success in invading genomes (Casacuberta and Santiago 2003; Feschotte and Pritham 2007). In accordance with this, it has been shown that the number of sequences related to the *Mariner*-like element *Lemi*1 is low in *Medicago truncatula*, where it has not given rise to MITEs, while it is much higher in *Arabidopsis* where it has given rise to the *Emigrant* MITE (Guermonprez et al. 2008).

7.4 Prevalence of MITEs and Their Impact in Plant Genomes

One of the characteristics that make MITEs a singular type of defective class II transposons is their capacity to reach high copy numbers in genomes (Casacuberta and Santiago 2003). MITEs are present in virtually all plant genomes, where their

copy number can vary but usually exceeds that of typical class II transposons. For example, more than 90,000 MITEs grouped into approximately 100 different families are present in the rice genome (Feschotte et al. 2003; Jiang et al. 2004; Juretic et al. 2004). Individual families such as the *Tourist* and S*towaway* families are found in more than 33,000 and 24,000 copies, respectively, in rice, and some 7,200 and 28,000 copies, respectively, in sorghum (Paterson et al. 2009). Even though these families are very large, the overall genome fraction MITEs occupy is relatively small, due to the diminutive size of these elements. Indeed, *Tourist* and *Stowaway* elements combined only occupy 3.24 % and 1.12 % of the rice and sorghum genomes, respectively, (Paterson et al. 2009). The size of a particular MITE family may vary greatly among closely related species and even between landraces. Indeed, it has been reported that while most rice strains only contain 1–50 copies of the *mPing* MITE, the EG4 strain and related landraces contain up to 1,000 (Naito et al. 2006). These data highlight these elements' capacity to multiply rapidly by bursts of amplification, which endows them with the capability to have an impact in genomes in spite of the low fraction they occupy.

Most MITEs are closely associated with genes in plant genomes. The first MITE described, *Tourist*, was shown to be closely associated to maize genes (Bureau and Wessler 1992), and this characteristic was found to be shared by most MITEs (Casacuberta and Santiago 2003; Wessler et al. 1995). For example, in rice and *Arabidopsis*, the majority of MITEs are located in the euchromatin (Feng et al. 2002; Santiago et al. 2002; Wright et al. 2003). This close association with genes could be the result of an insertion site preference or, alternatively, the effect of selection, as it seems to be the case for some *Arabidopsis* MITEs (Santiago et al. 2002). MITEs are not only located close to genes in plants but can also insert within genes, providing new promoter regulatory sequences (Naito et al. 2009; Sarilar et al. 2011), transcription termination elements (Kuang et al. 2009; Santiago et al. 2002), or even new alternative exons. Indeed, a recent report shows that the insertion of a MITE provides a functionally indispensable alternative exon in the tobacco mosaic virus N resistance gene (Kuang et al. 2009). While there are only a limited number of reports showing an unambiguous implication of MITE in creating new gene functions, there are many more examples of MITE insertions generating variability in gene sequences. A paradigmatic case is that of MITE insertions within resistance genes, which have been reported in rice (Song et al. 1998), barley (Wei et al. 2002), and potato (Huang et al. 2005). MITEs are also an important target of siRNAs, and their silencing may affect the expression of neighboring genes. The siRNAs that target MITEs can be of 24 nt (Kuang et al. 2009) or 21 nt (Cantu et al. 2010), suggesting that MITEs are targets of both transcriptional and posttranscriptional gene silencing. Thus, a MITE insertion within a gene promoter may attract heterochromatin and silence it transcriptionally, as it has been shown for other transposons (Lisch 2009), and an insertion within a transcribed region may make it prone to posttranscriptional gene silencing and mRNA degradation.

This close association with genes, together with their capability of reaching high copy numbers in short periods of time, makes MITEs a potent motor of gene evolution. MITE insertions polymorphic among accessions cultivars or lines have

been reported in pea, sugar beet, grapevine, potato, and *Medicago truncatul*a (Benjak et al. 2009; Grzebelus et al. 2009; Macas et al. 2005; Menzel et al. 2006; Momose et al. 2010), and occasionally this variability correlates with phenotypic differences (Momose et al. 2010). The analysis of a recent burst of amplification of the *mPing* element in rice shows an important number of insertions into the 5′ region of rice genes, which in some cases result in their transcriptional upregulation (Naito et al. 2009). The simultaneous insertion of different copies of the same MITE into different gene promoters may result in the coordinated regulation of multiple genes creating a so-called regulatory network (Feschotte 2008), as it has been proposed for *mPing* insertions in rice (Naito et al. 2009). However MITEs can also contribute to the coordinated expression of genes in a more subtle way. It has been shown that MITEs can encode miRNAs and siRNAs in plants (Kuang et al. 2009; Piriyapongsa and Jordan 2008). The frequent insertion of MITEs within transcribed regions of genes (Benjak et al. 2009; Kuang et al. 2009), and their capacity to form stable single strand secondary structures, may facilitate the production of siRNAs from the transcribed elements. Interestingly, it has been recently shown that MITE-derived siRNAs regulate ABA signaling and stress responses in rice (Yan et al. 2011). In this context, the insertion of multiple copies of the siRNA-producing MITE within different genes may also generate a regulatory network, as created by *mPing* MITE in rice (Naito et al. 2009).

7.5 Concluding Remarks

MITEs have been particularly successful in colonizing complex genomes. This is in part due to the difficulty of silencing them by homology-dependent pathways, as they are frequently mobilized by transposases to which they are only distantly related. Their success is probably also a consequence of their capacity to generate more subtle mutations than most other transposons. Indeed, MITEs are very short elements and their insertion within the non-translated regions of genes may be easier to tolerate. Their frequent association with genes, which seems more pronounced that that of their related DNA transposons, suggests that MITE insertions near or within genes have been selected for during evolution. The last few years have seen many reports highlighting the impact of these elements on plant genes' function and regulation, attesting to the role MITEs have played in the evolution of plant genomes.

References

Benjak A, Boue S, Forneck A, Casacuberta JM (2009) Recent amplification and impact of MITEs on the genome of grapevine (*Vitis vinifera* L.). Gen Biol Evol 1:75–84

Bergero R, Forrest A, Charlesworth D (2008) Active miniature transposons from a plant genome and its nonrecombining Y chromosome. Genetics 178:1085–1092

Bureau TE, Wessler SR (1992) Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4:1283–1294

Bureau TE, Wessler SR (1994) Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6:907–916

Bureau TE, Ronald PC, Wessler SR (1996) A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. Proc Natl Acad Sci USA 93:8524–8529

Cantu D, Vanzetti LS, Sumner A, Dubcovsky M, Matvienko M, Distelfeld A, Michelmore RW, Dubcovsky J (2010) Small RNAs, DNA methylation and transposable elements in wheat. BMC Genomics 11:408

Casacuberta JM, Santiago N (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene 311:1–11

Casacuberta E, Casacuberta JM, Puigdomenech P, Monfort A (1998) Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: characterisation of the Emigrant family of elements. Plant J 16:79–85

Chen Y, Zhou FF, Li GJ, Xu Y (2009) MUST: A system for identification of miniature inverted-repeat transposable elements and applications to *Anabaena variabilis* and *Haloquadratum walsbyi*. Gene 436:1–7

Deragon JM, Casacuberta JM, Panaud O (2008) Plant transposable elements. Genome Dyn 4:69–82

Dufresne M, Hua-Van A, El Wahab HA, Ben M'Barek S, Vasnier C, Teysset L, Kema GH, Daboussi MJ (2007) Transposition of a fungal miniature inverted-repeat transposable element through the action of a Tc1-like transposase. Genetics 175:441–452

Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X, Jia P, Zhao Q, Ying K, Yu S, Tang Y, Weng Q, Zhang L, Lu Y, Mu J, Zhang LS, Yu Z, Fan D, Liu X, Lu T, Li C, Wu Y, Sun T, Lei H, Li T, Hu H, Guan J, Wu M, Zhang R, Zhou B, Chen Z, Chen L, Jin Z, Wang R, Yin H, Cai Z, Ren S, Lv G, Gu W, Zhu G, Tu Y, Jia J, Chen J, Kang H, Chen X, Shao C, Sun Y, Hu Q, Zhang X, Zhang W, Wang L, Ding C, Sheng H, Gu J, Chen S, Ni L, Zhu F, Chen W, Lan L, Lai Y, Cheng Z, Gu M, Jiang J, Li J, Hong G, Xue Y, Han B (2002) Sequence and analysis of rice chromosome 4. Nature 420:316–320

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397–405

Feschotte C, Mouches C (2000) Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. Mol Biol Evol 17:730–737

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41:331–368

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3:329–341

Feschotte C, Swamy L, Wessler SR (2003) Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). Genetics 163:747–758

Feschotte C, Osterlund MT, Peeler R, Wessler SR (2005) DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. Nucleic Acids Res 33:2153–2165

Filee J, Siguier P, Chandler M (2007) Insertion sequence diversity in archaea. Microbiol Mol Biol Rev 71:121–157

Frith MC, Hamada M, Horton P (2010) Parameters for accurate genome alignment. BMC Bioinformatics 11:80

Grzebelus D, Gladysz M, Macko-Podgorni A, Gambin T, Golis B, Rakoczy R, Gambin A (2009) Population dynamics of miniature inverted-repeat transposable elements (MITEs) in *Medicago truncatula*. Gene 448:214–220

Guermonprez H, Loot C, Casacuberta JM (2008) Different strategies to persist: the pogo-like Lemi1 transposon produces miniature inverted-repeat transposable elements or typical defective elements in different plant genomes. Genetics 180:83–92

Guynet C, Hickman AB, Barabas O, Dyda F, Chandler M, Ton-Hoang B (2008) *In vitro* reconstitution of a single-stranded transposition mechanism of IS608. Mol Cell 29:302–312

Han Y, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38:e199

Han MJ, Shen YH, Gao YH, Chen LY, Xiang ZH, Zhang Z (2010) Burst expansion, distribution and diversification of MITEs in the silkworm genome. BMC Genomics 11:520

Hancock CN, Zhang F, Wessler SR (2010) Transposition of the Tourist-MITE mPing in yeast: an assay that retains key features of catalysis by the class 2 PIF/Harbinger superfamily. Mob DNA 1:5

Hancock CN, Zhang F, Floyd K, Richardson AO, Lafayette P, Tucker D, Wessler SR, Parrott WA (2011) The rice miniature inverted repeat transposable element mPing is an effective insertional mutagen in soybean. Plant Physiol 157:552–562

Huang S, van der Vossen E, Kuang H, Vleeshouwers V, Zhang N, Borm T, van Eck H, Baker B, Jacobsen E, Visser R (2005) Comparative genomics enabled the isolation of the R3a late blight resistance gene in potato. Plant J 42:251–261

Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, Wessler SR (2003) An active DNA transposon family in rice. Nature 421:163–167

Jiang N, Feschotte C, Zhang X, Wessler SR (2004) Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Curr Opin Plant Biol 7:115–119

Juretic N, Bureau TE, Bruskiewich RM (2004) Transposable element annotation of the rice genome. Bioinformatics 20:155–160

Kuang H, Padmanabhan C, Li F, Kamei A, Bhaskar PB, Ouyang S, Jiang J, Buell CR, Baker B (2009) Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: New functional implications for MITEs. Genome Res 19:42–56

Lisch D (2009) Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol 60:43–66

Loot C, Santiago N, Sanz A, Casacuberta JM (2006) The proteins encoded by the pogo-like Lemi1 element bind the TIRs and subterminal repeated motifs of the Arabidopsis Emigrant MITE: consequences for the transposition mechanism of MITEs. Nucleic Acids Res 34:5238–5246

Lyons M, Cardle L, Rostoks N, Waugh R, Flavell AJ (2008) Isolation, analysis and marker utility of novel miniature inverted repeat transposable elements from the barley genome. Mol Genet Genomics 280:275–285

Macas J, Koblizkova A, Neumann P (2005) Characterization of Stowaway MITEs in pea (*Pisum sativum* L.) and identification of their potential master elements. Genome 48:831–839

Mao L, Wood TC, Yu Y, Budiman MA, Tomkins J, Woo S, Sasinowski M, Presting G, Frisch D, Goff S, Dean RA, Wing RA (2000) Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res 10:982–990

Menzel G, Dechyeva D, Keller H, Lange C, Himmelbauer H, Schmidt T (2006) Mobilization and evolutionary history of miniature inverted-repeat transposable elements (MITEs) in *Beta vulgaris* L. Chromosome Res 14:831–844

Miskey C, Papp B, Mates L, Sinzelle L, Keller H, Izsvak Z, Ivics Z (2007) The ancient mariner sails again: transposition of the human Hsmar1 element by a reconstructed transposase and activities of the SETMAR protein on transposon ends. Mol Cell Biol 27:4589–4600

Momose M, Abe Y, Ozeki Y (2010) Miniature inverted-repeat transposable elements of Stowaway are active in potato. Genetics 186:59–66

Naito K, Cho E, Yang G, Campbell MA, Yano K, Okumoto Y, Tanisaka T, Wessler SR (2006) Dramatic amplification of a rice transposable element during recent domestication. Proc Natl Acad Sci USA 103:17620–17625

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461:1130–1134

Nakazaki T, Okumoto Y, Horibata A, Yamahira S, Teraishi M, Nishida H, Inoue H, Tanisaka T (2003) Mobilization of a transposon in the rice genome. Nature 421:170–172

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob ur R, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Petersen G, Seberg O (2000) Phylogenetic evidence for excision of Stowaway miniature inverted-repeat transposable elements in triticeae (Poaceae). Mol Biol Evol 17:1589–1596

Piriyapongsa J, Jordan IK (2007) A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS One 2:e203

Piriyapongsa J, Jordan IK (2008) Dual coding of siRNAs and miRNAs by plant transposable elements. RNA 14:814–821

Santiago N, Herraiz C, Goni JR, Messeguer X, Casacuberta JM (2002) Genome-wide analysis of the Emigrant family of MITEs of *Arabidopsis thaliana*. Mol Biol Evol 19:2285–2293

Sarilar V, Marmagne A, Brabant P, Joets J, Alix K (2011) BraSto, a Stowaway MITE from Brassica: recently active copies preferentially accumulate in the gene space. Plant Mol Biol 77:59–75

Schwarz-Sommer Z, Gubitz T, Weiss J, Gomez-di-Marco P, Delgado-Benarroch L, Hudson A, Egea-Cortines M (2010) A molecular recombination map of *Antirrhinum majus*. BMC Plant Biol 10:275

Sinzelle L, Jegot G, Brillet B, Rouleux-Bonnin F, Bigot Y, Auge-Gouillou C (2008) Factors acting on Mos1 transposition efficiency. BMC Mol Biol 9:106

Song WY, Pi LY, Bureau TE, Ronald PC (1998) Identification and characterization of 14 transposon-like elements in the noncoding regions of members of the Xa21 family of disease resistance genes in rice. Mol Gen Genet 258:449–456

Stratton M (2008) Genome resequencing and genetic variation. Nat Biotechnol 26:65–66

Surzycki SA, Belknap WR (2000) Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. Proc Natl Acad Sci USA 97:245–249

Ton-Hoang B, Pasternak C, Siguier P, Guynet C, Hickman AB, Dyda F, Sommer S, Chandler M (2010) Single-stranded DNA transposition is coupled to host replication. Cell 142:398–408

Tu ZJ (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. Proc Natl Acad Sci USA 98:1699–1704

Wei F, Wing RA, Wise RP (2002) Genome dynamics and evolution of the Mla (powdery mildew) resistance locus in barley. Plant Cell 14:1903–1917

Wessler SR, Bureau TE, White SE (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev 5:814–821

Wright SI, Agrawal N, Bureau TE (2003) Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. Genome Res 13:1897–1903

Yan Y, Zhang Y, Yang K, Sun Z, Fu Y, Chen X, Fang R (2011) Small RNAs from MITE-derived stem-loop precursors regulate abscisic acid signaling and abiotic stress responses in rice. Plant J 65:820–828

Yang G, Hall TC (2003a) MAK, a computational tool kit for automated MITE analysis. Nucleic Acids Res 31:3659–3665

Yang G, Hall TC (2003b) MDM-1 and MDM-2: two mutator-derived MITE families in rice. J Mol Evol 56:255–264

Yang G, Zhang F, Hancock CN, Wessler SR (2007) Transposition of the rice miniature inverted repeat transposable element mPing in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 104:10962–10967

Yang G, Nagel DH, Feschotte C, Hancock CN, Wessler SR (2009) Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. Science 325:1391–1394

Zhang X, Feschotte C, Zhang Q, Jiang N, Eggleston WB, Wessler SR (2001) P instability factor: an active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. Proc Natl Acad Sci USA 98:12572–12577